Using normative modelling to detect disease progression in mild cognitive impairment and Alzheimer's disease in a cross-sectional multi-cohort study

Walter H. L. Pinaya [a b c *], Cristina Scarpazza [a d], Rafael Garcia-Dias [a], Sandra Vieira [a], Lea Baecker [a], Pedro F. da Costa [e f], Alberto Redolfi [g], Giovanni B. Frisoni [h i], Michela Pievani [h], Vince D. Calhoun [j], João R. Sato [b], Andrea Mechelli [a]

* [a] Department of Psychosis Studies, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK.
[b] Center of Mathematics, Computing, and Cognition, Universidade Federal do ABC, Santo André, Brazil.
[c] Department of Biomedical Engineering, School of Biomedical Engineering & Imaging Sciences, King's College London, London, UK
[d] Department of General Psychology, University of Padua, Padua, Italy.
[e] Department of Neuroimaging, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK.
[f] Centre for Brain and Cognitive Development, Birkbeck College, University of London, London, UK.
[g] Laboratory of Neuroinformatics, IRCCS Istituto Centro San Giovanni di Dio Fatebenefratelli, Brescia, Italy.
[h] Laboratory of Alzheimer's Neuroimaging & Epidemiology, IRCCS Istituto Centro San Giovanni di Dio Fatebenefratelli, Brescia, Italy.
[i] Memory Clinic and LANVIE Laboratory of Neuroimaging of Aging, University Hospitals and University of Geneva, Geneva, Switzerland.
[j] Tri-institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS), Georgia State, Georgia Tech, Emory, US

*** Corresponding Author**: Walter H. L. Pinaya
Email address: walter.diaz_sanz@kcl.ac.uk

ABSTRACT

Normative modelling is an emerging method for quantifying how individuals deviate from the healthy populational pattern. Several machine learning models have been implemented to develop normative models to investigate brain disorders, including regression, support vector machines and Gaussian process models. With the advance of deep learning technology, the use of deep neural networks has also been proposed. In this study, we assessed normative models based on deep autoencoders using structural neuroimaging data from patients with Alzheimer's disease (n=206) and mild cognitive impairment (n=354). We first trained the autoencoder on an independent dataset (UK Biobank dataset) with 11,034 healthy controls. Then, we estimated how each patient deviated from this norm and established which brain regions were associated to this deviation. Finally, we compared the performance of our normative model against traditional classifiers. As expected, we found that patients exhibited deviations according to the severity of their clinical condition. The model identified medial temporal regions, including the hippocampus, and the ventricular system as critical regions for the calculation of the deviation score. Overall, the normative model had comparable cross-cohort generalizability to traditional classifiers. To promote open science, we are making all scripts and the trained models available to the wider research community.

Keywords: Normative modelling, neuroimaging, Alzheimer's disease, mild cognitive impairment, autoencoders, deep learning, anomaly detection

1. INTRODUCTION

Normative modelling is an emerging method for quantifying and describing how individuals deviate from the expected pattern learned from a population or large sample [1]. Recently, this approach has been applied to neuroimaging data to investigate a number of brain disorders, such as attention deficit hyperactivity disorder [2,3], autism spectrum disorder [4,5], schizophrenia [3,5,6] and dementia [7,8]. The procedure of normative modelling used in these studies has two steps: (i) first, statistical models are estimated to characterise the typical brain data from a reference cohort; (ii) then, the estimated model is applied to a target clinical cohort in order to quantify the variation (e.g. due to the effect of brain disorders).

Many statistical models have been proposed for normative modelling, including regression, support vector machines and Gaussian process modelling (for an extensive list, see Marquand et al., 2019). In Pinaya et al. [5], we proposed a normative modelling approach based on the use of deep autoencoders to evaluate psychiatric patients. The use of a deep learning approach [10,11] enables models to learn multiple levels of representation about the intricate structure of the data and identify the most important morphological characteristic of the healthy brain. In addition, in Pinaya et al. [5], the models were able to detect deviations at the level of the individual, with patients with schizophrenia and patients with autism spectrum disorder presenting values significantly higher than the healthy controls (HC).

Similar to psychiatric disorders, the clinical interpretation of magnetic resonance imaging scans can be challenging in the context of neurodegenerative disorders, as brain alterations may be difficult to distinguish from those related to healthy ageing. The identification of disease-related alterations can be

particularly tricky in the early stages of a disorder. For this reason, there is a grown interest in the development of methods for quantifying deviations of regional brain volumes that can discriminate between healthy and pathological ageing, with the ultimate aim of improving diagnostic and prognostic assessment of neurodegenerative disorders [12]. Here, we used the autoencoder normative method [5] to evaluate the most common type of dementia in the elderly worldwide, Alzheimer's disease (AD).

First, we trained the normative models using a large number of HC subjects (> 11,000 participants). Then, we assessed the performance of these models using data from patients with a diagnosis of mild cognitive impairment (MCI), the prodromal stage to AD, and patients with a diagnosis of AD. This assessment involved calculating the deviation, i.e. the extent to which subjects deviate from the norm, in five additional datasets composed of patients with MCI, patients with AD, and HC subjects. We had two main hypotheses. First, we hypothesised that the normative models would be robust and sensitive enough to create deviation values that reflect the severity of the brain anatomical alterations due to the disease, i.e. individuals with AD would deviate from normality more than those with MCI. Second, we hypothesised that the main brain regions driving the observed deviation would include the medial temporal cortex and the ventricular system, consistent with the results of previous neuroimaging studies of MCI and AD [13,14]. Finally, we compared the performance of the normative approach against traditional classifiers to discriminate the patient groups from the HC group.

2. METHODS

2.1. Datasets

In our analysis, we used six datasets: the UK Biobank [15], the Alzheimer's Disease Neuroimaging Initiative (ADNI) [16], the Australian Imaging Biomarkers and Lifestyle Study of Ageing (AIBL) [17], the Alzheimer's Disease Repository Without Borders (ARWiBo) [18,19], the Open Access Series of Imaging Studies: Cross-Sectional (OASIS-1) [20], and the Minimal Interval Resonance Imaging in Alzheimer's Disease (MIRIAD) [21].

The UK Biobank is a study that aims to follow the health and well-being of 500,000 volunteer participants across the United Kingdom. From these participants, a subsample was chosen to collect multimodal imaging, including structural neuroimaging. Here, we used an early release of the project's data comprising of 11,034 HC participants. The inclusion criteria for the present study were: a) subjects who had the data collected in the same MRI scanner (from Cheadle centre), b) age between 47 to 73 years old. The only exclusion criterion was previous hospitalization associated with the diagnosis of mental and behavioural disorders, disease of the nervous system, cerebrovascular diseases, benign neoplasm of meninges, brain and other parts of the central nervous system, or injuries to the head. This study (UK Biobank project #40323) was covered by the general ethical approval for UK Biobank studies from the NHS National Research Ethics Service on 17th June 2011 (Ref 11/NW/0382). All methods were carried out in accordance with the approved guidelines and regulations. All UK Biobank participants provided written informed consent. More details about the dataset can be found elsewhere [15,22–24].

The ADNI consortium started in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner. Its goal was to verify whether different neuroimaging biomarkers and neuropsychological assessments can be combined to measure the progression of MCI and to study the development of AD. All ADNI participants provided written informed consent, and study protocols were approved by each local site's institutional review board. All methods were carried out in accordance with the approved guidelines. Further information about ADNI, including full study protocols, complete inclusion and exclusion criteria, and data collection and availability can be found at http://www.adni-info.org/. All methods as stated on the website were performed with the relevant guidelines and regulations. In this study, we included the structural MRI collected during the ADNI GO, ADNI 2 and ADNI 3 phases. Similar to UK Biobank, we included only subjects with age between 47 to 73 years old. The final dataset comprised of 517 subjects, where 212 were HC, 159 were patients with early MCI (EMCI), 82 were patient with late MCI (LMCI), and 64 were patients with AD. In the ADNI datasets, participants were assigned to these MCI stages based on different levels of impairment on a single episodic memory measure, with the EMCI group showing milder episodic memory impairment than the LMCI group [25,26].

The AIBL dataset was developed to enhance the understanding of the pathogenesis of AD, concentrating on its early diagnosis (more details can be found in Ellis et al., 2009). Ethics approval for the AIBL study and all experimental protocols was provided by the ethics committees of Austin Health, St Vincent's Health, Hollywood Private Hospital and Edith Cowan University. All experiments and methods were carried out in accordance with the approved guidelines and regulations and all volunteers gave written informed consent before participating

in the study. Here, we included the structural MRI of subjects between 47 to 73 years old, to match the age range of the UK Biobank dataset. The final group was composed of 346 subjects, where 262 were HC, 46 were patients with MCI (stage not known), and 38 were patients with AD.

The ARWiBo is a cross-sectional dataset including data from patients and controls enrolled at the Scientific Institute for the Research and Care of Alzheimer's Disease [Istituto di Ricovero e Cura a Carattere Scientifico (IRCCS) Centro San Giovanni di Dio Fatebenefratelli, Brescia, Italy]. A multidisciplinary team of neurologists, neuroscientists, image analysists, neurophysiologists, and geneticists are involved in the assessment of patients. As part of their assessment, participants undergo blood drawing (for APOE genotyping), clinical and cognitive evaluations as well as high-resolution MRI scanning (more details can be found in Frisoni et al., 2009 and Galluzzi et al., 2010). Here, we included the structural MRI of subjects between 47 to 73 years old, to match the age range of the UK Biobank dataset. The resulting group was composed of 319 subjects, including 215 HC, 67 patients with MCI (stage not known), and 37 patients with AD. Ethics approval for the ARWiBo study and all experimental protocols was provided by the local ethics committee and all participants signed an informed participation consent. All experiments and methods were carried out in accordance with the approved guidelines and regulations.

The OASIS-1 dataset is the result of a collaborative effort of investigators from a single acquisition site supported by the National Institute on Aging (NIA), the Howard Hughes Medical Institute, the Biomedical Informatics Research Network (BIRN) and the Washington University Alzheimer's Disease Research Center [Alzheimer's Disease Research Center (ADRC)]. This collaborative effort

aimed to create a freely available MRI dataset for the wider scientific community. The original dataset consisted of a cross-sectional collection of subjects aged 18 to 96. It included participants over the age of 60 who had received a clinical diagnosis of very mild to moderate AD (for more information, please see http://www.oasis-brains.org). In our analysis, we selected data collected from individuals who were between 47 to 73 years old, to match the age range of the UK Biobank dataset. The resulting group was composed of 78 subjects, including 41 HC and 37 patients with AD. Ethics approval for the OASIS-1 study and all experimental protocols was provided by the local ethics committee and all participants signed an informed participation consent. All subjects participated in accordance with guidelines of the Washington University Human Studies Committee. All experiments and methods were carried out in accordance with the approved guidelines and regulations.

The MIRIAD dataset was designed to establish the minimal interval over which it would be feasible to undertake clinical trials in AD using atrophy measured from longitudinal MRI as an outcome measure [21]. Ethical approval for the MIRIAD study (and subsequently its release) was received from the local research ethics committee, and written consent obtained from all participants. All experiments and methods were carried out in accordance with the approved guidelines and regulations. Here, we included the structural MRI of subjects between 47 to 73 years old, to match the age range of the UK Biobank dataset. The resulting group was composed of 48 subjects, including 18 HC and 30 patients with AD.

In the present study, we used the UK Biobank set to train the autoencoders and the ADNI, AIBL, ARWiBo, OASIS-1, and MIRIAD datasets to assess the

normative model performance on data from patients with MCI and AD. To perform comparisons between HC and patient groups, we ensured that there were no significant statistical differences regarding age and sex in all three clinical datasets. We assessed each dataset independently using the ANOVA test to verify any differences in age and the Chi-square test of homogeneity to investigate differences in the sex ratios between groups (Table 1 and Table 2).

Table 1 - Demographic information for the subjects from the UK Biobank dataset, the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, and the Australian Imaging Biomarkers and Lifestyle Study of Ageing (AIBL) dataset. We used ANOVA test and the chi-square test of homogeneity to test for significant differences in age and sex between healthy controls and patients. Abbreviations: HC = healthy control; EMCI = early mild cognitive impairment; LMCI = late mild cognitive impairment; AD = Alzheimer's disease; MCI = mild cognitive impairment; SD = standard deviation.

	UK BIOBANK n=11,034	ADNI n=517				p	AIBL n=346			p
		HC n=212	EMCI n=159	LMCI n=82	AD n=64		HC n=262	MCI n=46	AD n=38	
Age, y						.87				.28
Mean ± SD	61.6±7.0	66.6±3.7	66.4±4.2	66.2±5.1	66.6±5.4		68.2±3.2	68.2±3.6	67.3±5.2	
Range	[47, 73]	[56, 72]	[56, 73]	[56, 73]	[56, 73]		[60, 73]	[56, 73]	[55, 73]	
Sex, n (%)						.25				.15
Men	5180 (47)	90 (42)	72 (45)	40 (49)	36 (56)		113 (43)	19 (41)	17 (45)	
Women	5854 (53)	122 (58)	87 (55)	42 (51)	28 (44)		149 (57)	27 (59)	21 (55)	

Table 2 - Demographic information for the subjects from the Alzheimer's Disease Repository Without Borders (ARWiBo) dataset, the Open Access Series of Imaging Studies: Cross-Sectional (OASIS-1) dataset, and the Minimal Interval Resonance Imaging in Alzheimer's Disease (MIRIAD) dataset. We used ANOVA test and the chi-square test of homogeneity to test for significant differences in age and sex between healthy controls and patients. Abbreviations: HC = healthy control; AD = Alzheimer's disease; MCI = mild cognitive impairment; SD = standard deviation.

	ARWiBo n=319			p	OASIS-1 N=78		p	MIRIAD n=48		p
	HC n=215	MCI n=67	AD n=37		HC n=41	AD n=37		HC n=18	AD n=30	
Age, y				.16			.07			.71
Mean ± SD	65.1±4.4	66.4±5.7	65.1±6.0		68.2±3.8	69.7±3.1		66.7±4.1	66.2±4.6	
Range	[57, 73]	[47, 73]	[50, 73]		[61, 73]	[62, 73]		[59, 73]	[56, 73]	
Sex, n (%)				.70			.30			.76
Men	86 (40)	23 (34)	14 (38)		11 (27)	14 (38)		7 (39)	13 (43)	
Women	129 (60)	44 (66)	23 (62)		30 (73)	23 (62)		11 (61)	17 (57)	

2.2. MRI Processing

We used the FreeSurfer software (version 6.0) to estimate the brain regions' volumes from the T1 weighted images. This estimation was performed using the "recon-all" command (see Fischl, 2012; Fischl et al., 2002, for more information). During this processing, the cortical surface of each hemisphere was parcellated according to the Desikan-Killiany atlas [29] and anatomical volumetric measures were obtained via a whole-brain segmentation procedure (Aseg atlas) [28]. The final data included the cortical volume for each of the 68 cortical subregions (34 per hemisphere) and the volume of 33 neuroanatomical structures, totalling 101 subregions/structures (the complete list is presented in the supplementary materials).

2.3. Normative model

In this paper, we developed the normative model using the adversarial autoencoder (AAE; Figure 1) [30,31]. As an autoencoder, this neural network has an encoder and a decoder. The function of the encoder is to take in an input **x** and map it into a latent encoding space, creating a latent code **h**. Then, the goal of the decoder is to reconstruct the input data based on the latent code. The AAE is a blend of this autoencoder framework with adversarial training, which is used in generative adversarial networks modelling [32]. This autoencoder uses the adversarial training to shape the distribution of the latent code to look similar to a predefined prior distribution. The AAE achieves this desired distribution by incorporating a discriminator network into its structure. In this scheme, the discriminator receives two types of inputs: random numbers sampled from the

desired prior distribution, and the latent code. During the training process, the discriminator will make predictions regarding whether its input data was sampled from the prior distribution or the latent code. The adversarial training forces the encoder to produce a latent code space that can fool the discriminator into predicting that the encoded samples are just another sample from the prior distribution.

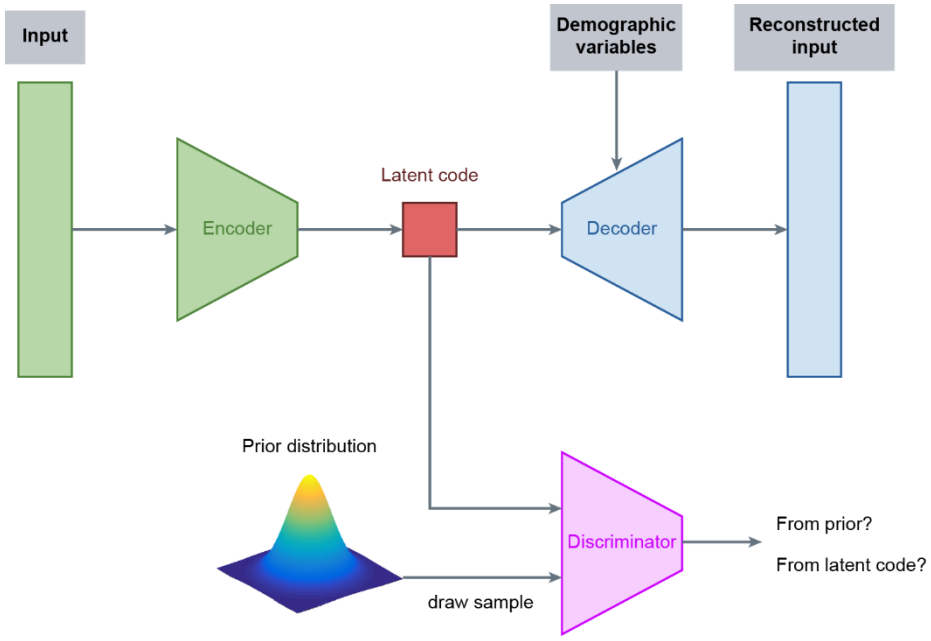


Figure 1 – Structure of the normative model based on adversarial autoencoders. In this configuration, the subject data is inputted into the encoder and then mapped to the latent code. This latent code is fed to the decoder with the demographic data, and then the decoder generates a reconstruction of the original data. During the training of the model, the discriminator predicts if its input data came from the latent code or if it was randomly sampled from the chosen prior distribution (e.g. Gaussian distribution). Based on these predictions, the adversarial autoencoder forces the encoder to produce a latent code similar to the prior distribution selected. Since the model is trained on healthy controls data, it is expected that it can reconstruct similar data relatively well, yielding a small reconstruction error. However, the model is expected to generate a high error when processing data affected by unseen underlying mechanisms, e.g. pathological mechanisms.

In this study, we trained the AAE to codify and reconstruct the data of HC subjects. The main idea of this normative approach is that, since the AAE only

learns how to reconstruct images from HC individuals, it will be less precise at mapping images from patients, which differ due to the pathological mechanisms of the disorder. As a result, the difference between the reconstructed data and the original data will be larger in patients than HC individuals.

Regarding our model architecture, we used an encoder with two hidden layers with 100 neurons, and a latent code with a size of 20 neurons. The decoder and the discriminator had a similar structure (two hidden layers with 100 neurons). All hidden layers had a leaky ReLU non-linearity [33]. The latent code and the decoder's output layer had a linear activation function.

2.4. Normative model training

To train the autoencoder, first, we performed the pre-processing of the brain features. This involved estimating the relative brain region volumes for each subject by dividing the original brain region volumes by the total intracranial volume. Then, we normalised the relative brain region volumes across all the participants in the training set. In this step, we performed a normalisation robust to outliers by subtracting the median value of the relative brain region volume and then scaling the data according to its interquartile range. Centering and scaling was done independently for each brain region. The same relevant statistics (median and interquartile range) were later used to normalise the data from the clinical datasets before feeding them to the model.

In our analyses, we used a conditioned AAE [30]. This type of autoencoder allows us to influence the model's reconstruction using the demographic variables, i.e. age and sex. To input these variables into the model, we transformed age and sex into one-hot encoding vectors. After this transformation,

each subject has an age vector with 27 positions, where each position corresponds to a year within the range of 47-73 years. In this vector, all positions have value zero except the one that indicates the subject's age which has a value equal to 1. The subject's sex was represented in a one-hot encoded vector with two positions, one for male and one for female. The AAE's decoder used these vectors together with the latent code to reconstruct the brain data. This architecture forces the network to disentangle the label information from the latent code [30].

With the features pre-processed and the conditioning data prepared, we trained the autoencoder to minimise the mean squared value of its reconstruction error using Adam optimizer [34] for 200 epochs. A minibatch approach was used in this gradient descent-based optimizer, with a batch size of 256. The model was trained with a cyclical learning rate [35], which allows convergence of the training with fewer epochs. We started using a base learning rate with a value of 0.0001 and a maximum learning rate value of 0.005, chosen using the "LR Range Test" [36]. The learning rate cycle had a basic triangular shape with an amplitude decaying (gamma = 0.98).

In this study, we accessed the robustness of the autoencoder approach by training it with different simulated sets using the bootstrapping as the resampling method. We created 1,000 bootstrapped sets (each one with n = 11,032) by sampling with replacement from the UK Biobank. These bootstrapped sets were used to train the AAE. With this resampling method, we calculated: the value of the mean deviation (section 2.5) for each group from the ADNI, AIBL, ARWiBo, OASIS-1, and MIRIAD datasets, the discriminative performance of the normative

approach (section 2.5), and the deviation from normality of each brain region (section 2.6).

2.5. Analysis of the observed deviations

Similar to Pinaya et al. [5], we processed the data of each subject using the AAE, and we calculated the mean squared error between the reconstruction and the inputted data as the metric of brain deviation (Eq. 1).

$$observed\ deviation = \frac{1}{number\ of\ regions} \sum_{i=1}^{number\ of\ regions} (x_i - \hat{x}_i)^2 \qquad (Eq.\ 1)$$

where x_i is the normalised value of the brain region i, $\hat{x}_i$ is the autoencoder reconstructed value of the brain region i, and $number\ of\ regions$ is the number of cortical regions and neuroanatomical structures used (i.e. $number\ of\ regions$ = 101).

In each iteration of the bootstrap method, we used the trained autoencoder to obtain the deviation metric of the subjects from the ADNI, AIBL, ARWiBo, OASIS-1, and MIRIAD datasets. Then, we calculated the difference between the mean deviation scores of each pair of groups. We identified a significant difference between groups if the confidence interval (95% of confidence) of this difference did not include the zero. Besides, we used the subjects' deviations to obtain the discriminative performance of the autoencoder approach, measured by the area under the receiver operating characteristic curve (AUC).

2.6. Brain regions deviations

The autoencoder approach can quantify how much each brain region deviated from normality and contributed to the observed deviation. These values

were obtained by measuring the difference between the inputted value and its reconstruction. In our study, we quantified the deviation for each subject from the ADNI, AIBL, ARWiBo, OASIS-1, and MIRIAD datasets. Then, in each iteration of the bootstrap method, we calculated the effect size of each brain region deviation – using Cliff's delta [37] value - between the HC group and each patient group. Here we used Cliff's delta - a non-parametric effect size measure – because the observed deviation presents a gamma distribution.

2.7. Comparison against traditional machine learning classification

A further aim of the present study was to compare the performance of our normative model against a traditional classification approach. To measure the performance of the classifiers, we calculated the AUC using the .632+ bootstrap method [38] with 1,000 iterations. Each clinical dataset (ADNI, AIBL, ARWiBo, OASIS-1, and MIRIAD) was analysed independently using the HC and patient groups to train the classifiers. Besides, the analysis was performed as multiple binary classifications between HC and each clinical group (e.g. HC versus LMCI).

In each iteration, first, we created the bootstrapped set by sampling the original data (from ADNI, AIBL, ARWiBo, OASIS-1, and MIRIAD datasets) with replacement. This bootstrapped set had the same size as the original dataset (for example, when analysing the ADNI dataset to classify healthy controls and patients with Alzheimer's disease, the bootstrapped set had 212+64=276 subjects), and it contained repeated subjects (due to replacement). For each iteration, the subjects not included in the bootstrapped set were used as the out-of-bag set (i.e. test set).

Next, we obtained the relative brain region volumes of each subject by dividing the original volume by the total intracranial volume. Then, we normalised

the values of the relative brain volumes across the subjects. In this normalisation step, we removed the median value of the brain regions and scaled the data according to the interquartile range. Centering and scaling was done independently for each brain region. The same relevant statistics (median and interquartile range) were later used to normalise the out-of-bag set.

To perform the classification analysis, we used a relevance vector machine (RVM) [39] with a linear kernel. The RVM is a Bayesian treatment of identical functional form to the Support Vector Machines (SVM) [40]. One advantage of the RVM form over the SVM is that it is not necessary to estimate the error/margin trade-off parameter 'C'. After we trained the RVM on the bootstrapped set, we used the model to obtain the predicted probability of a subject belonging to the patient class. Using these probabilities, we calculated two AUC values, one for the bootstrapped set (called "resubstitution" metric) and one for the test set (called "out-of-bag" metric). By using the .632+ bootstrap method, we minimised the optimistic and pessimistic bias of the estimate and obtained the AUC value (Eq.2).

$$AUC_{bootstrap} = \frac{1}{b}\sum_{i=1}^{b}\left(\omega * AUC_{out-of-bag,i} + (1-\omega) * AUC_{resubstitution,i}\right) \quad (\mathrm{Eq.}\,2)$$

where b was the number of iterations and the weight ω was defined considering the relative overfitting rate (full description in Efron and Tibshirani, 1997). To obtain the confidence interval (CI; 95% of confidence), we used the percentile method [41]. Next, we compared these confidence intervals with the AUC obtained during the normative approach.

Finally, we compared the generalization of the classifiers with the results of the autoencoders. In this analysis, we used each trained classifier to predict

the group of the subjects from the other clinical datasets. In order to verify if the performance in the independent datasets was significantly different from the normative approach, we calculated the difference between the AUCs of this generalization analysis and the AUCs of the autoencoders. With the 1,000 measures of the difference, we calculated its confidence interval (95% confidence) to verify if this difference is different from zero.

2.8. Experiments

We conducted our experiments in Python 3 using the Tensorflow 2.0 library (https://www.tensorflow.org/) and the sklearn_rvm library (https://github.com/Mind-the-Pineapple/sklearn-rvm) developed by Baecker et al. [42]. We have made publicly available the codes and trained models used in this study at https://github.com/Warvito/Normative-modelling-using-deep-autoencoders. A Google's Colaboratory notebook that calculates the deviations scores of new data is available at https://colab.research.google.com/github/Warvito/Normative-modelling-using-deep-autoencoders/blob/master/notebooks/predict.ipynb.

3. RESULTS

3.1. Comparison of deviation values for healthy controls and patients

Figure 2 shows the mean value of the observed deviation for each group. For the ADNI dataset, we found a mean value of 0.28 ([0.27, 0.32]; 95% CI) for HC; 0.29 ([0.28, 0.35]; 95% CI) for EMCI; 0.32 ([0.30, 0.38]; 95% CI) for LMCI; 0.37 ([0.34, 0.47]; 95% CI) for AD. For the AIBL dataset, we found a mean value of 0.30 ([0.28, 0.33]; 95% CI) for HC; 0.36 ([0.34, 0.42]; 95% CI) for MCI; and 0.40 ([0.36, 0.50]; 95% CI) for AD. For the ARWiBo dataset, we found a mean value of 0.32 ([0.30, 0.38]; 95% CI) for HC; 0.37 ([0.34, 0.47]; 95% CI) for MCI; and 0.46 ([0.40, 0.62]; 95% CI) for AD. For the OASIS-1 dataset, we found a mean value of 0.41 ([0.39, 0.46]; 95% CI) for HC and 0.65 ([0.58, 0.79]; 95% CI) for AD. For the MIRIAD dataset, we found a mean value of 0.26 ([0.24, 0.30]; 95% CI) for HC and 0.48 ([0.41, 0.71]; 95% CI) for AD.

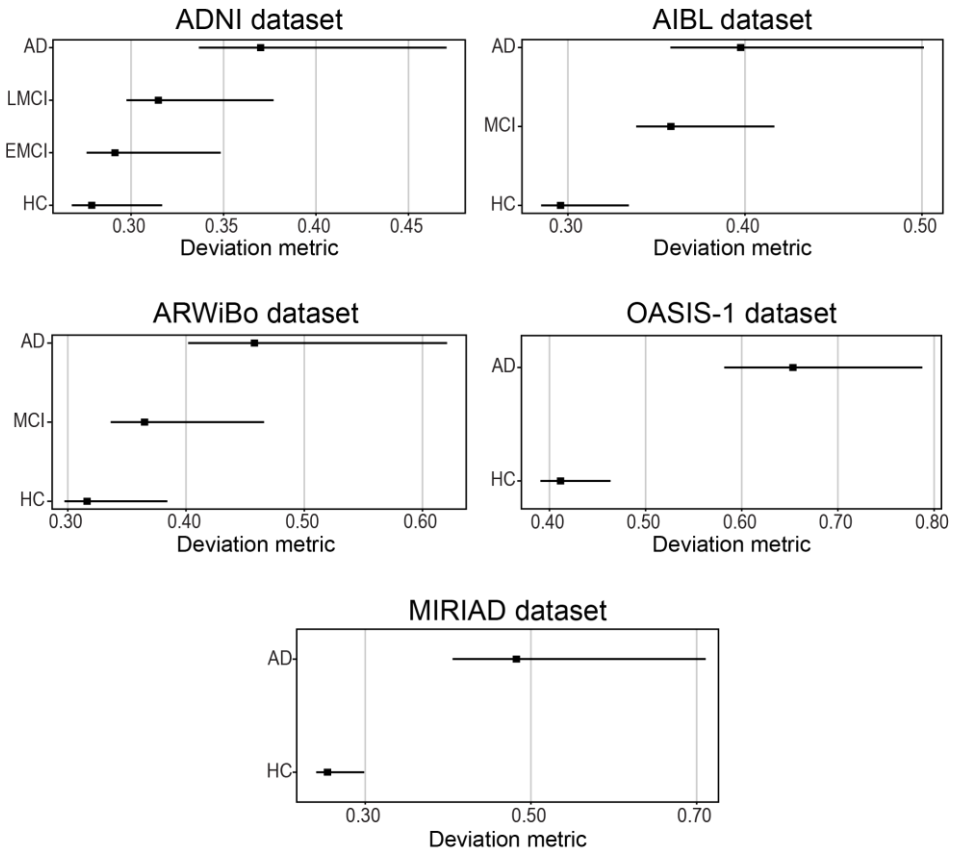


Figure 2 – Mean value of the observed deviation calculated by the autoencoder for each group. The square marker indicates the mean value and the horizontal bars indicates the 95% confidence interval calculated using the percentile method on the bootstrap analysis. Abbreviations: AD = Alzheimer's disease; EMCI = early mild cognitive impairment; LMCI = late mild cognitive impairment; MCI = mild cognitive impairment; HC = healthy controls; ADNI = Alzheimer's Disease Neuroimaging Initiative; AIBL = Australian Imaging Biomarkers and Lifestyle Study of Ageing; ARWiBo = Alzheimer's Disease Repository Without Borders; OASIS-1 = Open Access Series of Imaging Studies: Cross-Sectional; MIRIAD = Minimal Interval Resonance Imaging in Alzheimer's Disease.

When we examined the confidence intervals of the observed deviations, we found that the five independent datasets presented mean deviation scores significantly different between groups, with the exception of the comparison between HC and EMCI in the ADNI dataset (difference range [-0.03, 0.00]) and the comparison between MCI and AD in the AIBL dataset (difference range [-0.09, 0.00]) (more details can be found in the supplementary materials).

3.2. Normative model performance in discriminative tasks

We examined if the observed deviations could be used to predict if a person belonged to the patient or HC group (Figure 3) using ROC curves. This revealed that the generated deviation values reflected the severity of the disease. Specifically, based on the AUC, it was possible to discriminate patients with AD vs HC better than patients with MCI vs HC, and to discriminate patients with LMCI vs HC better than patients with EMCI vs HC.

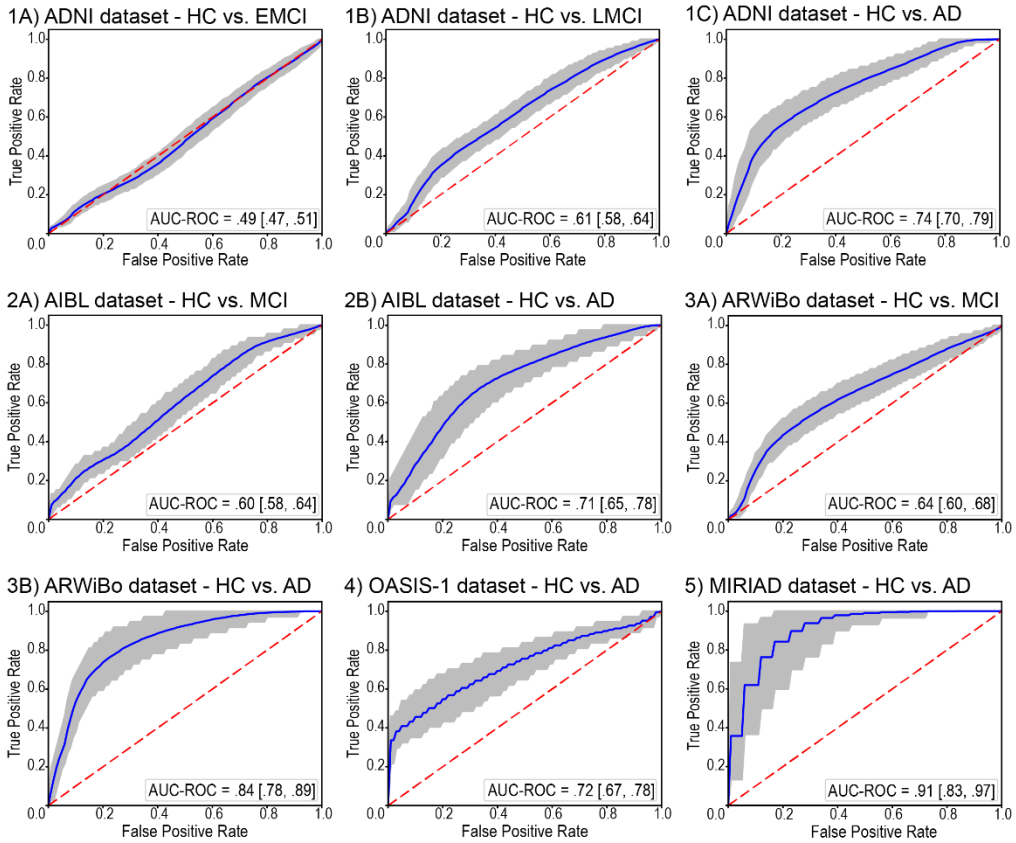


Figure 3 – Discriminative performance of the normative approach. The solid line indicates the mean receiver operating characteristic curve across the bootstrap iterations with the shaded area indicating the 95% confidence interval calculated using the percentile method on the bootstrap analysis. The dashed line indicates the chance level. Abbreviations: AD = Alzheimer's disease; AUC-ROC = area under the receiver operating characteristic curve; EMCI = early mild cognitive impairment; LMCI = late mild cognitive impairment; MCI = mild cognitive impairment; HC = healthy controls; ADNI = Alzheimer's Disease Neuroimaging Initiative; AIBL = Australian Imaging Biomarkers and Lifestyle Study of Ageing; ARWiBo = Alzheimer's Disease Repository Without Borders; OASIS-1 = Open Access Series of Imaging Studies: Cross-Sectional; MIRIAD = Minimal Interval Resonance Imaging in Alzheimer's Disease.

3.3. Brain regions deviations

Figure 4 present the Cliff's delta of each brain region when comparing its deviation in the HC group against the deviation in the patient groups. Only the regions with effect sizes significantly different from zero are shown (complete list presented in the supplementary materials). Among the regions showing significant deviation in patients with AD, we found the lateral ventricles, temporal horns, hippocampus, entorhinal cortex, parahippocampal cortex, and amygdala. A number of these regions also showed a high deviation in patients with MCI, including the lateral ventricles and hippocampus. Finally, we also noted that effect sizes were smaller for the regions identified in patients with MCI relative to those identified in patients with AD.

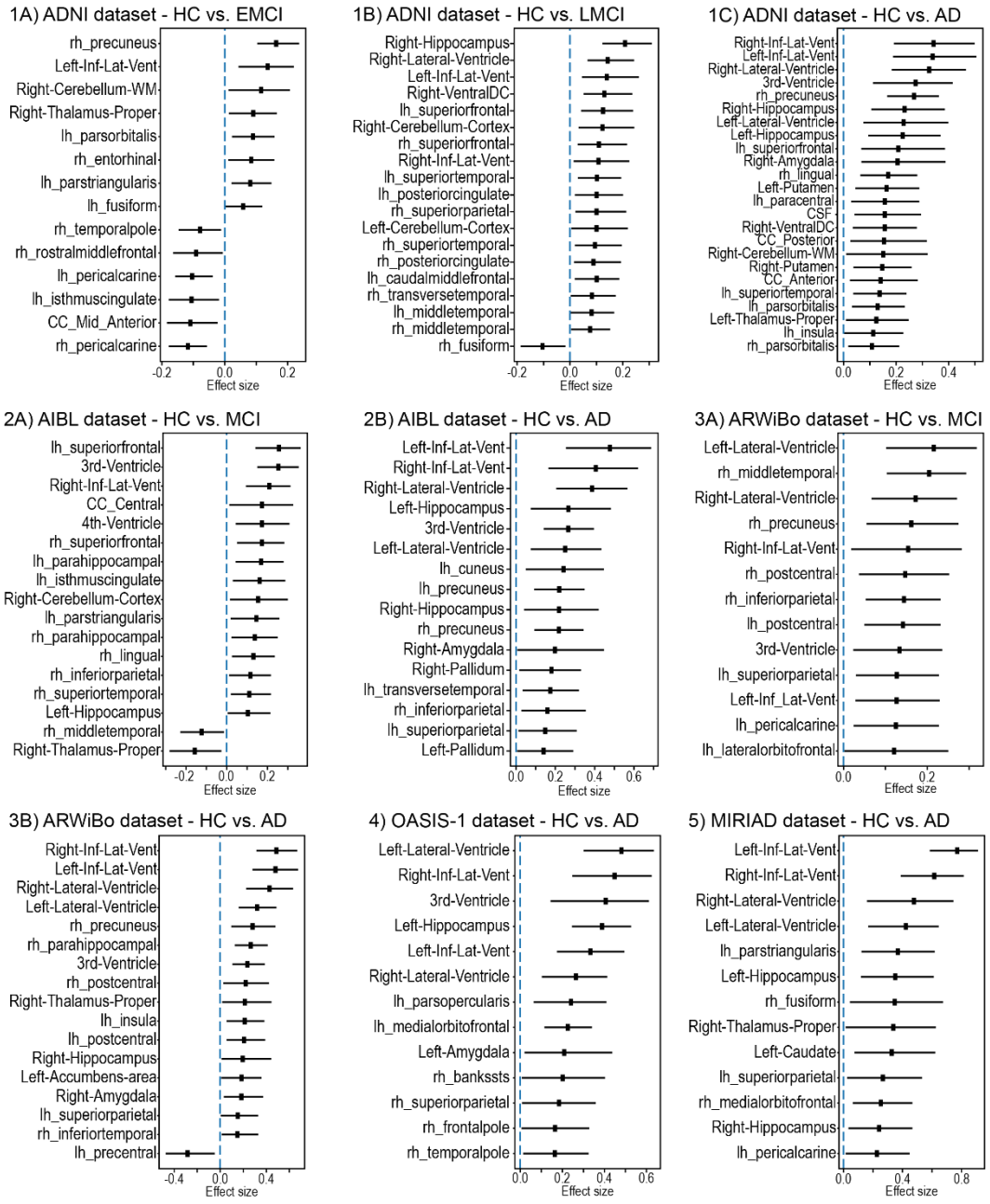


Figure 4 – Brain regions deviations. The square marker indicates the mean effect size (Cliff's delta) between the healthy control group and the respective patient groups. The horizontal bars indicate the 95% confidence interval calculated using the percentile method on the bootstrap analysis. Only the regions with a mean effect size significantly different from zero are presented. Abbreviations: AD = Alzheimer's disease; AUC-ROC = area under the receiver operating characteristic curve; EMCI = early mild cognitive impairment; LMCI = late mild cognitive impairment; MCI = mild cognitive impairment; HC = healthy controls; ADNI = Alzheimer's Disease Neuroimaging Initiative; AIBL = Australian Imaging Biomarkers and Lifestyle Study of Ageing; ARWiBo = Alzheimer's Disease Repository Without Borders; OASIS-1 = Open Access Series of Imaging Studies: Cross-Sectional; MIRIAD = Minimal Interval Resonance Imaging in Alzheimer's Disease.

3.4. Traditional machine learning classification

Using the RVM, we verified the performance of a traditional classifier when performing binary classification between HC and patients. For the ADNI dataset, we obtained an AUC = 0.69 ([0.58, 0.77]; 95% CI) when analysing patients with EMCI, an AUC = 0.76 ([0.64, 0.84]; 95% CI) when analysing patients with LMCI, and an AUC = 0.93 ([0.87, 0.97]; 95% CI) when analysing patients with AD. For the AIBL dataset, an AUC = 0.37 ([0.00, 0.78]; 95% CI) when analysing subjects with MCI, and we obtained an AUC = 0.93 ([0.86, 0.93]; 95% CI) when analysing patients with AD. Note, that the AUC for the AIBL dataset when analysing MCI had a wide interval. This interval was exacerbated due to the presence of overfitting and the .632+ bootstrap method compensatory effect that reduce the effect of bias caused by this overfitting. For the ARWiBo dataset, we obtained an AUC = 0.68 ([0.52, 0.78]; 95% CI) when analysing subjects with MCI, and an AUC = 0.94 ([0.87, 0.98]; 95% CI) when analysing patients with AD. For the OASIS-1 dataset, we obtained an AUC = 0.86 ([0.69, 0.96]; 95% CI) when analysing patients with AD. For the MIRIAD dataset, we obtained an AUC = 0.86 ([0.70, 0.96]; 95% CI) when analysing patients with AD.

To identify significant differences between the performance of the normative models and traditional classifiers, we calculated the confidence interval (95% of confidence) of the difference in AUC between the two methods. The traditional classifiers were superior to the normative models when predicting the difference between the groups in the ADNI dataset and the difference between HC and AD in the AIBL dataset; in contrast the performance of the two approaches was comparable for all other comparisons (more details can be found in the supplementary materials).

Finally, we examined how a classifier trained on a certain dataset would perform when applied to other datasets (i.e. cross-cohort generalizability). The results of this examination are presented in Table 3 and Table 4. When predicting AD, the classifiers had a higher mean performance than the normative approach in most cases (except when the model was trained on MIRIAD dataset and evaluated on ARWiBo dataset). However, the difference was not significantly different in almost half of the cases. When predicting MCI, the classifiers presented a lower mean performance in all cases, but the difference was not significantly different.

Table 3 – Generalization performance of the classifiers for the classification between HC and patients with Alzheimer's disease. In this table, the rows indicate the dataset where the classifier is trained and the columns indicate the dataset where the performance was tested. The area under the receiver operating characteristic curve is shown with the upper and lower bound of its 95% confidence interval. Performance significantly different from the normative approach calculated using the confidence interval of the difference between the approach across the bootstrap scheme is indicated by "*".

	ADNI	AIBL	ARWiBo	OASIS-1	MIRIAD
ADNI	-	0.89 [0.93, 0.83] *	0.88 [0.81, 0.93]	0.84 [0.76, 0.90] *	0.98 [0.95, 1.00]
AIBL	0.88 [0.82, 0.93] *	-	0.89 [0.93, 0.83] *	0.86 [0.80, 0.91] *	0.98 [0.94, 1.00]
ARWiBo	0.88 [0.81, 0.92] *	0.88 [0.83, 0.92] *	-	0.80 [0.72, 0.86]	0.96 [0.91, 0.99]
OASIS-1	0.90 [0.84, 0.93] *	0.89 [0.83, 0.93] *	0.86 [0.77, 0.92]	-	0.96 [0.91, 1.00]
MIRIAD	0.89 [0.83, 0.93] *	0.87 [0.80, 0.91] *	0.82 [0.73, 0.91]	0.83 [0.74, 0.89]	-

Table 4 - Generalization performance of the classifiers for the classification between HC and patients with mild cognitive impairment. In this table, the rows indicate the dataset where the classifier is trained and the columns indicate the dataset where the performance was measured. The area under the receiver operating characteristic curve is shown with the upper and lower bound of its 95% confidence interval. No case had a performance significantly different from the normative approach calculated using the confidence interval of the difference between the approach across the bootstrap scheme.

	AIBL	ARWiBo
AIBL	-	0.61 [0.54, 0.67]
ARWiBo	0.59 [0.53, 0.65]	-

4. DISCUSSION

In this study, we evaluated the performance of the normative modelling approach based on deep autoencoders on data from patients with MCI and AD. Consistent with our first hypothesis, we found that the approach was effective in generating deviation values that reflect the severity of the disease, with patients with AD showing higher deviations than patients with MCI, and patients with LMCI showing larger deviations than patients with EMCI. We also measured how much each brain region deviated from normality and contributed to the observed deviation. Here, we found that regions from the ventricular system and medial temporal lobe were among those making the greatest significant contribution to deviation, consistent with our second hypothesis. Finally, we compared the performance of the normative approach versus a traditional classification approach. Although a higher performance was found for traditional classifiers in most cases, the difference was not statistically significant in the majority of cases.

We have replicated previous findings that the autoencoder is capable of detecting neuroanatomical deviation in individuals with brain disorders [5]. In particular, in each of our five independent datasets, the normative model was able to assign higher values to patients with AD than healthy controls. This pattern was expected since the disorder is associated with profound alterations in the brain morphometry which were not present in the training set [13,14]. In addition, we have expanded these findings by demonstrating for the first time that autoencoders are capable of discriminating between different stages of the disease progress (i.e. EMCI versus LMCI versus AD). In particular, we observed that the MCI group presented intermediary deviation values in three independent datasets (ADNI, AIBL and ARWiBo). These values were also expected since the

MCI is considered as a transitory stage between HC and AD [43], and usually present less brain atrophy compared to AD [44]. In addition, within the ADNI dataset, the MCI subjects were divided into two categories, EMCI and LMCI. Although individuals in both stages meet the conventional criteria for MCI, EMCI is associated with less pronounced symptoms thought to reflect an earlier point in the clinical spectrum than LMCI. In our analyses, we found that the patients with LMCI had a significantly (i.e. the confidence interval of the difference between the group do not overlap zero) larger deviation than patients with EMCI providing further confirmation that that deep autoencoders are capable of discriminating between different stages of the disease course.

With the autoencoder based approach, it was possible to identify the brain regions with the highest deviations from the expected normative values. Consistent with our second hypothesis, the AD group showed high levels of deviation in structures that are part of the ventricular system (such as the lateral ventricles, temporal horns, and 3rd ventricle) and in the medial temporal cortex, including the hippocampus, entorhinal cortex, parahippocampal cortex, and amygdala. Progressive ventricular expansion is one of the most reliable morphological changes in dementia patients, reflecting the increasing atrophy of the brain [45]. Likewise, medial temporal cortex atrophy is among the most consistent findings in neuroimaging studies of AD [13,46] and an established marker of AD [47]. While deviations in the MCI group had a smaller sizes than those in the AD group, there was a high degree of overlap in the hippocampus, parahippocampal cortex and several temporoparietal regions, consistent with previous neuroimaging studies of MCI [48–51]. The smaller effect size in MCI might be explained by two (not mutually-exclusive) factors: i) earlier stage in the AD

course, hence milder atrophy, ii) heterogeneity of the MCI construct. Since MCI patients were not selected based on AD biomarkers (i.e., presence of beta-amyloid and tau protein in the cerebrospinal fluid) [52], this group will likely include a mixture of AD and non-AD cases, hence the milder/diluted effect.

Finally, we compared the performance of our normative approach with traditional classifiers. The performance of the classifiers was measured in two schemes, on data from the same dataset where the model was trained and on data from independent clinical datasets (generalization performance). Although the traditional classifiers had a better mean performance in most cases, the differences between the two approaches were not statistically significant in most of the cases, especially when predicting the subjects from the ARWiBo, OASIS-1 and MIRIAD datasets. This similarity was more evident during the prediction of the patients with MCI (with exception the ADNI dataset).

Although we evaluated our method using a range of different datasets, we did not assess the impact of MRI scanners and acquisition parameters. Recent studies have showed that these variables can have a measurable impact on the performance of machine learning models, highlighting the importance of inter-scanner harmonisation [53,54]. In particular, MRI scanners and acquisition parameters have been shown to influence the results not only in traditional machine learning classification but also normative modelling [55]. For this reason, further studies need to be performed to analyse the influence of inter-scanner harmonisation, which can be implemented using tools such as Neuroharmony [54] or Combat [56], on the performance of autoencoder based methods.

Different from a case-control context, the normative approach does not need to be trained in a dataset with reasonable balancing between HC and

patient groups. It is trained using only healthy controls, which enables the use of large cohorts of HC participants [1,9], such as UK Biobank and Human Connectome Project [57]. Our approach is not linked with any labels during training; this enables its application to an array of clinical tasks (including diagnosis, prognosis, treatment selection and mechanistic inference) for any brain disorder without the necessity of re-training or fine-tuning. Finally, since our approach involves anomaly detection, it can also work cooperatively with conventional discriminative models to identify and mitigate circumstances where supervised methods could catastrophic fail due to a test example very distinct from the training set ("out-of-distribution" examples). In order to promote open science, we have made all scripts and the trained models available to the wider research community (https://github.com/Warvito/Normative-modelling-using-deep-autoencoders).

ACKNOWLEDGEMENTS

This study was supported by a Wellcome Trust's Innovator Award to Andrea Mechelli (208519/Z/17/Z). WHLP is supported by Wellcome Innovations [WT213038/Z/18/Z]. This work was carried out within the scope of the project "use-inspired basic research", for which the Department of General Psychology of the University of Padova has been recognized as "Dipartimento di eccellenza" by the Ministry of University and Research. JRS was supported by grant 2018/04654-9, São Paulo Research Foundation (FAPESP); and grant 2018/21934-5, São Paulo Research Foundation (FAPESP). VC was supported by NIH RF1AG063153.

This research has been conducted using the UK Biobank Resource (Project number: 40323).

Data were provided in part by OASIS: Cross-Sectional: Principal Investigators: D. Marcus, R, Buckner, J, Csernansky J. Morris; P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, U24 RR021382.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La

Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

Data used in the preparation of this article was obtained from the Australian Imaging Biomarkers and Lifestyle flagship study of ageing (AIBL) funded by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) which was made available at the ADNI database (www.loni.usc.edu/ADNI). The AIBL researchers contributed data but did not participate in analysis or writing of this report. AIBL researchers are listed at www.aibl.csiro.au.

ARWiBo data (www.arwibo.it) was obtained from NeuGRID4You initiative funded by the European Commission (FP7/2007-2013) under grant agreement no.283562. The overall goal of ARWiBo is to contribute, thorough synergy with neuGRID (https://neugrid2.eu), to global data sharing and analysis in order to develop effective therapies, prevention methods and a cure for Alzheimer' and other neurodegenerative diseases.

Data used in the preparation of this article were obtained from the MIRIAD database. The MIRIAD investigators did not participate in analysis or writing of this report. The MIRIAD dataset is made available through the support of the UK Alzheimer's Society (Grant RF116). The original data collection was funded through an unrestricted educational grant from GlaxoSmithKline (Grant 6GKC).

AUTHOR CONTRIBUTIONS

WHLP obtained and organized the MRI data, pre-processed the MRI images, implemented the normative models, performed the groups analysis, drafted and edited the manuscript; CS obtained and organized the MRI data and revised the manuscript; RGD implemented script to help pre-process the MRI data and the revised the manuscript; SV revised the manuscript; LB revised the manuscript; PFC implemented the traditional classifiers and revised the manuscript; AR was responsible for part of the data collection and revised the manuscript; MP was responsible for part of the data collection and revised the manuscript; GBF was responsible for part of the data collection; VDC revised manuscript and involved in funding efforts; JRS revised manuscript and involved in funding efforts; AM revised the manuscript, gathered funding, and supervised the project.

Competing interests

The author(s) declare no competing interests.

Data availability

UK Biobank data are available through a procedure described at http://www.ukbiobank.ac.uk/using-the-resource/. ADNI data are available through an access procedure described at http://adni.loni.usc.edu/data-samples/access-data/. AIBL data are available through an access procedure described at https://aibl.csiro.au/adni/imaging.html. AIBL data are available through an access procedure described at https://www.gaaindata.org/partner/ARWIBO. OASIS-1 data are available through an access procedure described at https://www.oasis-brains.org/#data. OASIS-1 data are available through an access procedure described at https://www.oasis-brains.org/#data. MIRIAD data are available through an access procedure described at https://www.ucl.ac.uk/drc/research/methods/minimal-interval-resonance-imaging-alzheimers-disease-miriad.

REFERENCES

1. Marquand, A. F., Rezek, I., Buitelaar, J. & Beckmann, C. F. Understanding heterogeneity in clinical cohorts using normative models: beyond case-control studies. *Biol. Psychiatry* **80**, 552–561 (2016).

2. Wolfers, T. *et al.* Individual differences v. the average patient: mapping the heterogeneity in ADHD using normative models. *Psychol. Med.* 1–10 (2019).

3. Kia, S. M. & Marquand, A. F. Neural Processes Mixed-Effect Models for Deep Normative Modeling of Clinical Neuroimaging Data. *arXiv Prepr. arXiv1812.04998* (2018).

4. Zabihi, M. *et al.* Dissecting the heterogeneous cortical anatomy of autism spectrum disorder using normative models. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **4**, 567–578 (2019).

5. Pinaya, W. H. L., Mechelli, A. & Sato, J. R. Using deep autoencoders to identify abnormal brain structural patterns in neuropsychiatric disorders: A large-scale multi-sample study. *Hum. Brain Mapp.* **40**, 944–954 (2019).

6. Wolfers, T. *et al.* Mapping the Heterogeneous Phenotype of Schizophrenia and Bipolar Disorder Using Normative Models. *JAMA Psychiatry* 1–10 (2018) doi:10.1001/JAMAPSYCHIATRY.2018.2467.

7. Ziegler, G., Ridgway, G. R., Dahnke, R., Gaser, C. & Initiative, A. D. N. Individualized Gaussian process-based prediction and detection of local and global gray matter abnormalities in elderly subjects. *Neuroimage* **97**, 333–348 (2014).

8. Huizinga, W. *et al.* A spatio-temporal reference model of the aging brain. *Neuroimage* **169**, 11–22 (2018).

9. Marquand, A. F. *et al.* Conceptualizing mental disorders as deviations from normative functioning. *Mol. Psychiatry* 1 (2019).

10. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436 (2015).

11. Vieira, S., Pinaya, W. H. L., Garcia-Dias, R. & Mechelli, A. Deep neural networks. in *Machine Learning* 157–172 (Elsevier, 2020).

12. Brewer, J. B. Fully-automated volumetric MRI with normative ranges: translation to clinical practice. *Behav. Neurol.* **21**, 21–28 (2009).

13. Busatto, G. F., Diniz, B. S. & Zanetti, M. V. Voxel-based morphometry in Alzheimer's disease. *Expert Rev. Neurother.* **8**, 1691–1702 (2008).

14. Pini, L. *et al.* Brain atrophy in Alzheimer's disease and aging. *Ageing Res. Rev.* **30**, 25–48 (2016).

15. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).

16. Mueller, S. G. *et al.* Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimer's Dement.* **1**, 55–66 (2005).

17. Ellis, K. A. *et al.* The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. *Int. Psychogeriatrics* **21**, 672–687 (2009).

18. Frisoni, G. B. *et al.* Markers of Alzheimer's disease in a population attending a memory clinic. *Alzheimer's Dement.* **5**, 307–317 (2009).

19. Galluzzi, S. *et al.* The new Alzheimer's criteria in a naturalistic series of patients with mild cognitive impairment. *J. Neurol.* **257**, 2004–2014 (2010).

20. Marcus, D. S. *et al.* Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *J. Cogn. Neurosci.* **19**, 1498–1507 (2007).

21. Malone, I. B. *et al.* MIRIAD—Public release of a multiple time point Alzheimer's MR imaging dataset. *Neuroimage* **70**, 33–36 (2013).

22. Elliott, P. & Peakman, T. C. The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. *Int. J. Epidemiol.* **37**, 234–244 (2008).

23. Miller, K. L. *et al.* Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat. Neurosci.* **19**, 1523 (2016).

24. Alfaro-Almagro, F. *et al.* Image processing and quality control for the first 10,000 brain imaging datasets from UK Biobank. *Neuroimage* **166**, 400–424 (2018).

25. Aisen, P. S. *et al.* Clinical Core of the Alzheimer's Disease Neuroimaging Initiative: progress and plans. *Alzheimer's Dement.* **6**, 239–246 (2010).

26. Edmonds, E. C. *et al.* Early versus late MCI: Improved MCI staging using a neuropsychological approach. *Alzheimer's Dement.* **15**, 699–708 (2019).

27. Fischl, B. FreeSurfer. *Neuroimage* **62**, 774–781 (2012).

28. Fischl, B. *et al.* Whole Brain Segmentation. *Neuron* **33**, 341–355 (2002).

29. Desikan, R. S. *et al.* An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* **31**, 968–980 (2006).

30. Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I. & Frey, B. Adversarial autoencoders. *arXiv Prepr. arXiv1511.05644* (2015).

31. Pinaya, W. H. L., Vieira, S., Garcia-Dias, R. & Mechelli, A. Autoencoders. in *Machine Learning* 193–208 (Elsevier, 2020).

32. Goodfellow, I. *et al.* Generative adversarial nets. in *Advances in neural information processing systems* 2672–2680 (2014).

33. Maas, A. L., Hannun, A. Y. & Ng, A. Y. Rectifier nonlinearities improve neural network acoustic models. in *Proc. icml* vol. 30 3 (2013).

34. Kingma, D. & Ba, J. Adam: A method for stochastic optimization. *arXiv Prepr. arXiv1412.6980* 1–15 (2014).

35. Smith, L. N. Cyclical learning rates for training neural networks. in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)* 464–472 (IEEE, 2017).

36. Smith, L. N. A disciplined approach to neural network hyper-parameters: Part 1--learning rate, batch size, momentum, and weight decay. *arXiv Prepr. arXiv1803.09820* (2018).

37. Cliff, N. Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychol. Bull.* **114**, 494 (1993).

38. Efron, B. & Tibshirani, R. Improvements on cross-validation: the 632+ bootstrap method. *J. Am. Stat. Assoc.* **92**, 548–560 (1997).

39. Tipping, M. E. The relevance vector machine. in *Advances in neural information processing systems* 652–658 (2000).

40. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).

41. Efron, B. Nonparametric standard errors and confidence intervals. *Can. J. Stat.* **9**, 139–158 (1981).

42. Baecker, L. *et al.* Brain age prediction: A comparison between machine learning models using region- and voxel-based morphometric data. *Hum. Brain Mapp.* (2021).

43. Morris, J. C. *et al.* Mild cognitive impairment represents early-stage Alzheimer disease. *Arch. Neurol.* **58**, 397–405 (2001).

44. Pihlajamaki, M., Jauhiainen, A. M. & Soininen, H. Structural and functional MRI in mild cognitive impairment. *Curr. Alzheimer Res.* **6**, 179–185 (2009).

45. Thompson, P. M. *et al.* Mapping hippocampal and ventricular change in Alzheimer disease. *Neuroimage* **22**, 1754–1766 (2004).

46. Fox, N. C. & Schott, J. M. Imaging cerebral atrophy: normal ageing to Alzheimer's disease. *Lancet* **363**, 392–394 (2004).

47. Drago, V. *et al.* Disease tracking markers for Alzheimer's disease at the prodromal (MCI) stage. *J. Alzheimer's Dis.* **26**, 159–199 (2011).

48. Chételat, G. *et al.* Mapping gray matter loss with voxel-based morphometry in mild cognitive impairment. *Neuroreport* **13**, 1939–1943 (2002).

49. Hämäläinen, A. *et al.* Voxel-based morphometry to detect brain atrophy in progressive mild cognitive impairment. *Neuroimage* **37**, 1122–1131 (2007).

50. Pennanen, C. *et al.* A voxel based morphometry study on mild cognitive impairment. *J. Neurol. Neurosurg. Psychiatry* **76**, 11–14 (2005).

51. Kang, D. W., Lim, H. K., Joo, S., Lee, N. R. & Lee, C. U. Differential associations between volumes of atrophic cortical brain regions and memory performances in early and late mild cognitive impairment. *Front. Aging Neurosci.* **11**, 245 (2019).

52. Mulder, C. *et al.* Amyloid-β (1–42), total tau, and phosphorylated tau as cerebrospinal fluid biomarkers for the diagnosis of Alzheimer disease. *Clin. Chem.* **56**, 248–253 (2010).

53. Fortin, J. P. *et al.* Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage* **167**, 104–120 (2018).

54. Garcia-Dias, R. *et al.* Neuroharmony: A new tool for harmonizing volumetric MRI data from unseen scanners. *Neuroimage* **220**, (2020).

55. Kia, S. M. *et al.* Federated Multi-Site Normative Modeling using Hierarchical Bayesian Regression. *bioRxiv* (2021).

56. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).

57. Van Essen, D. C. *et al.* The WU-Minn Human Connectome Project: An overview. *Neuroimage* **80**, 62–79 (2013).

Using normative modelling to detect disease progression in mild cognitive impairment and Alzheimer's disease in a cross-sectional multi-cohort study

Walter H. L. Pinaya [a b c *], Cristina Scarpazza [a d], Rafael Garcia-Dias [a], Sandra Vieira [a], Lea Baecker [a], Pedro F. da Costa [e f], Alberto Redolfi [g], Giovanni B. Frisoni [h i], Michela Pievani [h], Vince D. Calhoun [j], João R. Sato [b], Andrea Mechelli [a]

* [a] Department of Psychosis Studies, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK.
[b] Center of Mathematics, Computing, and Cognition, Universidade Federal do ABC, Santo André, Brazil.
[c] Department of Biomedical Engineering, School of Biomedical Engineering & Imaging Sciences, King's College London, London, UK
[d] Department of General Psychology, University of Padua, Padua, Italy.
[e] Department of Neuroimaging, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK.
[f] Centre for Brain and Cognitive Development, Birkbeck College, University of London, London, UK.
[g] Laboratory of Neuroinformatics, IRCCS Istituto Centro San Giovanni di Dio Fatebenefratelli, Brescia, Italy.
[h] Laboratory of Alzheimer's Neuroimaging & Epidemiology, IRCCS Istituto Centro San Giovanni di Dio Fatebenefratelli, Brescia, Italy.
[i] Memory Clinic and LANVIE Laboratory of Neuroimaging of Aging, University Hospitals and University of Geneva, Geneva, Switzerland.
[j] Tri-institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS), Georgia State, Georgia Tech, Emory, US

1. List of brain features

Left-Lateral-Ventricle	lh_parstriangularis_volume
Left-Inf-Lat-Vent	lh_pericalcarine_volume
Left-Cerebellum-White-Matter	lh_postcentral_volume
Left-Cerebellum-Cortex	lh_posteriorcingulate_volume
Left-Thalamus-Proper	lh_precentral_volume
Left-Caudate	lh_precuneus_volume
Left-Putamen	lh_rostralanteriorcingulate_volume
Left-Pallidum	lh_rostralmiddlefrontal_volume
3rd-Ventricle	lh_superiorfrontal_volume
4th-Ventricle	lh_superiorparietal_volume
Brain-Stem	lh_superiortemporal_volume
Left-Hippocampus	lh_supramarginal_volume
Left-Amygdala	lh_frontalpole_volume
CSF	lh_temporalpole_volume
Left-Accumbens-area	lh_transversetemporal_volume
Left-VentralDC	lh_insula_volume
Right-Lateral-Ventricle	rh_bankssts_volume
Right-Inf-Lat-Vent	rh_caudalanteriorcingulate_volume
Right-Cerebellum-White-Matter	rh_caudalmiddlefrontal_volume
Right-Cerebellum-Cortex	rh_cuneus_volume
Right-Thalamus-Proper	rh_entorhinal_volume
Right-Caudate	rh_fusiform_volume
Right-Putamen	rh_inferiorparietal_volume
Right-Pallidum	rh_inferiortemporal_volume
Right-Hippocampus	rh_isthmuscingulate_volume
Right-Amygdala	rh_lateraloccipital_volume
Right-Accumbens-area	rh_lateralorbitofrontal_volume
Right-VentralDC	rh_lingual_volume
CC_Posterior	rh_medialorbitofrontal_volume
CC_Mid_Posterior	rh_middletemporal_volume
CC_Central	rh_parahippocampal_volume
CC_Mid_Anterior	rh_paracentral_volume
CC_Anterior	rh_parsopercularis_volume
lh_bankssts_volume	rh_parsorbitalis_volume
lh_caudalanteriorcingulate_volume	rh_parstriangularis_volume
lh_caudalmiddlefrontal_volume	rh_pericalcarine_volume
lh_cuneus_volume	rh_postcentral_volume
lh_entorhinal_volume	rh_posteriorcingulate_volume
lh_fusiform_volume	rh_precentral_volume
lh_inferiorparietal_volume	rh_precuneus_volume
lh_inferiortemporal_volume	rh_rostralanteriorcingulate_volume
lh_isthmuscingulate_volume	rh_rostralmiddlefrontal_volume
lh_lateraloccipital_volume	rh_superiorfrontal_volume
lh_lateralorbitofrontal_volume	rh_superiorparietal_volume
lh_lingual_volume	rh_superiortemporal_volume
lh_medialorbitofrontal_volume	rh_supramarginal_volume
lh_middletemporal_volume	rh_frontalpole_volume
lh_parahippocampal_volume	rh_temporalpole_volume
lh_paracentral_volume	rh_transversetemporal_volume
lh_parsopercularis_volume	rh_insula_volume
lh_parsorbitalis_volume	

2. Univariate analysis – ADNI dataset– HC vs EMCI

Supplementary Table 1 - Statistical significance measured by the Mann-Whitney U test and effect size measured by Cliff's delta absolute value based on the comparison of the reconstruction error for each brain region between the HC and the EMCI groups from the ADNI dataset. The regions with p-value <= 0.05 are highlighted in bold.

Regions	Effect size	p-value	Regions	Effect size	p-value
Left-Lateral-Ventricle	-0.045	0.230	**lh_parstriangularis_volume**	**0.130**	**0.016**
Left-Inf-Lat-Vent	**-0.112**	**0.033**	lh_pericalcarine_volume	0.046	0.226
Left-Cerebellum-White-Matter	-0.039	0.260	lh_postcentral_volume	0.077	0.101
Left-Cerebellum-Cortex	0.094	0.060	**lh_posteriorcingulate_volume**	**0.140**	**0.010**
Left-Thalamus-Proper	**0.209**	**>0.001**	lh_precentral_volume	0.060	0.160
Left-Caudate	0.062	0.153	**lh_precuneus_volume**	**0.085**	**0.081**
Left-Putamen	0.025	0.339	lh_rostralanteriorcingulate_volume	0.095	0.059
Left-Pallidum	-0.019	0.380	**lh_rostralmiddlefrontal_volume**	**0.124**	**0.021**
3rd-Ventricle	-0.021	0.365	**lh_superiorfrontal_volume**	**0.148**	**0.007**
4th-Ventricle	**0.153**	**0.006**	lh_superiorparietal_volume	0.047	0.220
Brain-Stem	0.090	0.069	**lh_superiortemporal_volume**	**0.139**	**0.011**
Left-Hippocampus	0.085	0.082	lh_supramarginal_volume	0.034	0.289
Left-Amygdala	0.079	0.097	lh_frontalpole_volume	-0.071	0.121
CSF	**-0.108**	**0.038**	**lh_temporalpole_volume**	**0.143**	**0.009**
Left-Accumbens-area	0.008	0.447	lh_transversetemporal_volume	0.033	0.295
Left-VentralDC	0.038	0.265	**lh_insula_volume**	**0.189**	**0.001**
Right-Lateral-Ventricle	-0.050	0.205	rh_bankssts_volume	0.005	0.466
Right-Inf-Lat-Vent	**-0.103**	**0.046**	rh_caudalanteriorcingulate_volume	0.006	0.460
Right-Cerebellum-White-Matter	-0.065	0.142	**rh_caudalmiddlefrontal_volume**	**0.132**	**0.015**
Right-Cerebellum-Cortex	**0.112**	**0.033**	rh_cuneus_volume	-0.031	0.306
Right-Thalamus-Proper	**0.199**	**0.001**	rh_entorhinal_volume	0.010	0.437
Right-Caudate	0.072	0.116	rh_fusiform_volume	0.062	0.153
Right-Putamen	0.010	0.437	rh_inferiorparietal_volume	0.065	0.141
Right-Pallidum	-0.016	0.398	rh_inferiortemporal_volume	0.024	0.345
Right-Hippocampus	**0.123**	**0.022**	rh_isthmuscingulate_volume	0.068	0.130
Right-Amygdala	0.008	0.448	rh_lateraloccipital_volume	-0.012	0.420
Right-Accumbens-area	**-0.107**	**0.039**	**rh_lateralorbitofrontal_volume**	**0.091**	**0.067**
Right-VentralDC	0.027	0.328	rh_lingual_volume	-0.050	0.207
CC_Posterior	-0.035	0.284	**rh_medialorbitofrontal_volume**	**0.156**	**0.005**
CC_Mid_Posterior	0.029	0.318	**rh_middletemporal_volume**	**0.100**	**0.050**
CC_Central	**0.103**	**0.045**	rh_parahippocampal_volume	0.029	0.315
CC_Mid_Anterior	0.031	0.304	**rh_paracentral_volume**	**0.113**	**0.032**
CC_Anterior	-0.033	0.292	**rh_parsopercularis_volume**	**0.109**	**0.036**
lh_bankssts_volume	**0.118**	**0.026**	rh_parsorbitalis_volume	0.025	0.338
lh_caudalanteriorcingulate_volume	0.065	0.140	rh_parstriangularis_volume	0.075	0.109
lh_caudalmiddlefrontal_volume	0.031	0.306	rh_pericalcarine_volume	0.078	0.099
lh_cuneus_volume	-0.041	0.248	**rh_postcentral_volume**	**0.101**	**0.048**
lh_entorhinal_volume	**-0.092**	**0.065**	rh_posteriorcingulate_volume	0.056	0.179
lh_fusiform_volume	0.052	0.195	**rh_precentral_volume**	**0.105**	**0.041**
lh_inferiorparietal_volume	**0.100**	**0.049**	rh_precuneus_volume	0.042	0.242
lh_inferiortemporal_volume	**0.092**	**0.064**	**rh_rostralanteriorcingulate_volume**	**0.112**	**0.033**
lh_isthmuscingulate_volume	-0.048	0.213	rh_rostralmiddlefrontal_volume	0.084	0.083
lh_lateraloccipital_volume	-0.002	0.484	**rh_superiorfrontal_volume**	**0.114**	**0.030**
lh_lateralorbitofrontal_volume	0.077	0.101	rh_superiorparietal_volume	0.030	0.308
lh_lingual_volume	0.007	0.457	**rh_superiortemporal_volume**	**0.221**	**0.000**
lh_medialorbitofrontal_volume	**0.232**	**0.000**	rh_supramarginal_volume	0.083	0.084
lh_middletemporal_volume	**0.109**	**0.037**	**rh_frontalpole_volume**	**-0.104**	**0.043**
lh_parahippocampal_volume	-0.014	0.406	**rh_temporalpole_volume**	**0.116**	**0.028**
lh_paracentral_volume	**0.126**	**0.019**	**rh_transversetemporal_volume**	**0.139**	**0.011**
lh_parsopercularis_volume	**0.134**	**0.014**	**rh_insula_volume**	**0.172**	**0.002**
lh_parsorbitalis_volume	-0.012	0.424			

3. Univariate analysis – ADNI dataset – HC vs LMCI

Supplementary Table 2 - Statistical significance measured by the Mann-Whitney U test and effect size measured by Cliff's delta absolute value based on the comparison of the reconstruction error for each brain region between the HC and the LMCI groups from the ADNI dataset. The regions with p-value <= 0.05 are highlighted in bold.

Regions	Effect size	p-value	Regions	Effect size	p-value
Left-Lateral-Ventricle	**-0.106**	**0.080**	**lh_parstriangularis_volume**	**0.232**	**0.001**
Left-Inf-Lat-Vent	**-0.299**	**>0.001**	lh_pericalcarine_volume	0.065	0.193
Left-Cerebellum-White-Matter	-0.054	0.238	**lh_postcentral_volume**	**0.191**	**0.006**
Left-Cerebellum-Cortex	**0.127**	**0.045**	**lh_posteriorcingulate_volume**	**0.132**	**0.040**
Left-Thalamus-Proper	**0.240**	**0.001**	**lh_precentral_volume**	**0.147**	**0.026**
Left-Caudate	0.098	0.096	**lh_precuneus_volume**	**0.246**	**0.001**
Left-Putamen	0.098	0.096	lh_rostralanteriorcingulate_volume	0.113	0.066
Left-Pallidum	0.016	0.415	**lh_rostralmiddlefrontal_volume**	**0.239**	**0.001**
3rd-Ventricle	**-0.161**	**0.016**	**lh_superiorfrontal_volume**	**0.287**	**>0.001**
4th-Ventricle	0.039	0.301	lh_superiorparietal_volume	0.120	0.055
Brain-Stem	0.068	0.183	**lh_superiortemporal_volume**	**0.298**	**>0.001**
Left-Hippocampus	**0.366**	**>0.001**	**lh_supramarginal_volume**	**0.231**	**0.001**
Left-Amygdala	**0.278**	**>0.001**	lh_frontalpole_volume	0.123	0.051
CSF	**-0.245**	**0.001**	**lh_temporalpole_volume**	**0.142**	**0.029**
Left-Accumbens-area	**0.222**	**0.002**	lh_transversetemporal_volume	0.086	0.126
Left-VentralDC	0.112	0.068	**lh_insula_volume**	**0.254**	**>0.001**
Right-Lateral-Ventricle	-0.106	0.079	**rh_bankssts_volume**	**0.227**	**0.001**
Right-Inf-Lat-Vent	**-0.258**	**>0.001**	rh_caudalanteriorcingulate_volume	0.001	0.493
Right-Cerebellum-White-Matter	-0.108	0.076	**rh_caudalmiddlefrontal_volume**	**0.142**	**0.029**
Right-Cerebellum-Cortex	0.118	0.058	rh_cuneus_volume	0.079	0.147
Right-Thalamus-Proper	**0.336**	**>0.001**	**rh_entorhinal_volume**	**0.214**	**0.002**
Right-Caudate	0.066	0.192	**rh_fusiform_volume**	**0.198**	**0.004**
Right-Putamen	0.087	0.125	**rh_inferiorparietal_volume**	**0.345**	**>0.001**
Right-Pallidum	-0.044	0.280	**rh_inferiortemporal_volume**	**0.267**	**>0.001**
Right-Hippocampus	**0.382**	**>0.001**	rh_isthmuscingulate_volume	0.106	0.080
Right-Amygdala	**0.157**	**0.018**	**rh_lateraloccipital_volume**	**0.173**	**0.011**
Right-Accumbens-area	0.026	0.363	**rh_lateralorbitofrontal_volume**	**0.125**	**0.048**
Right-VentralDC	0.106	0.080	rh_lingual_volume	0.081	0.141
CC_Posterior	0.013	0.429	**rh_medialorbitofrontal_volume**	**0.189**	**0.006**
CC_Mid_Posterior	-0.032	0.337	**rh_middletemporal_volume**	**0.349**	**>0.001**
CC_Central	0.083	0.134	**rh_parahippocampal_volume**	**0.178**	**0.009**
CC_Mid_Anterior	0.105	0.081	**rh_paracentral_volume**	**0.167**	**0.013**
CC_Anterior	0.012	0.436	**rh_parsopercularis_volume**	**0.126**	**0.047**
lh_bankssts_volume	**0.335**	**>0.001**	**rh_parsorbitalis_volume**	**0.139**	**0.032**
lh_caudalanteriorcingulate_volume	-0.092	0.110	**rh_parstriangularis_volume**	**0.212**	**0.002**
lh_caudalmiddlefrontal_volume	**0.240**	**0.001**	rh_pericalcarine_volume	0.052	0.245
lh_cuneus_volume	0.012	0.435	rh_postcentral_volume	0.089	0.120
lh_entorhinal_volume	**0.210**	**0.003**	**rh_posteriorcingulate_volume**	**0.265**	**>0.001**
lh_fusiform_volume	**0.371**	**>0.001**	rh_precentral_volume	0.034	0.327
lh_inferiorparietal_volume	**0.377**	**>0.001**	**rh_precuneus_volume**	**0.237**	**0.001**
lh_inferiortemporal_volume	**0.227**	**0.001**	rh_rostralanteriorcingulate_volume	0.057	0.226
lh_isthmuscingulate_volume	**0.130**	**0.043**	**rh_rostralmiddlefrontal_volume**	**0.271**	**>0.001**
lh_lateraloccipital_volume	0.122	0.053	**rh_superiorfrontal_volume**	**0.249**	**>0.001**
lh_lateralorbitofrontal_volume	**0.145**	**0.027**	**rh_superiorparietal_volume**	**0.241**	**0.001**
lh_lingual_volume	0.112	0.069	**rh_superiortemporal_volume**	**0.334**	**>0.001**
lh_medialorbitofrontal_volume	**0.197**	**0.004**	**rh_supramarginal_volume**	**0.230**	**0.001**
lh_middletemporal_volume	**0.344**	**>0.001**	rh_frontalpole_volume	-0.029	0.352
lh_parahippocampal_volume	**0.167**	**0.013**	rh_temporalpole_volume	0.089	0.118
lh_paracentral_volume	**0.125**	**0.048**	rh_transversetemporal_volume	0.033	0.332
lh_parsopercularis_volume	**0.247**	**0.001**	**rh_insula_volume**	**0.204**	**0.003**
lh_parsorbitalis_volume	0.075	0.160			

4. Univariate analysis – ADNI dataset – HC vs AD

Supplementary Table 3 - Statistical significance measured by the Mann-Whitney U test and effect size measured by Cliff's delta absolute value based on the comparison of the reconstruction error for each brain region between the HC and the AD groups from the ADNI dataset. The regions with p-value <= 0.05 are highlighted in bold.

Regions	Effect size	p-value	Regions	Effect size	p-value
Left-Lateral-Ventricle	**-0.552**	**>0.001**	**lh_parstriangularis_volume**	**0.420**	**>0.001**
Left-Inf-Lat-Vent	**-0.708**	**>0.001**	**lh_pericalcarine_volume**	**0.246**	**0.001**
Left-Cerebellum-White-Matter	-0.040	0.316	**lh_postcentral_volume**	**0.330**	**>0.001**
Left-Cerebellum-Cortex	0.103	0.106	**lh_posteriorcingulate_volume**	**0.414**	**>0.001**
Left-Thalamus-Proper	**0.488**	**>0.001**	**lh_precentral_volume**	**0.317**	**>0.001**
Left-Caudate	**0.179**	**0.015**	**lh_precuneus_volume**	**0.640**	**>0.001**
Left-Putamen	**0.378**	**>0.001**	**lh_rostralanteriorcingulate_volume**	**0.328**	**>0.001**
Left-Pallidum	-0.052	0.264	**lh_rostralmiddlefrontal_volume**	**0.581**	**>0.001**
3rd-Ventricle	**-0.542**	**>0.001**	**lh_superiorfrontal_volume**	**0.633**	**>0.001**
4th-Ventricle	-0.075	0.181	**lh_superiorparietal_volume**	**0.424**	**>0.001**
Brain-Stem	**0.189**	**0.011**	**lh_superiortemporal_volume**	**0.676**	**>0.001**
Left-Hippocampus	**0.748**	**>0.001**	**lh_supramarginal_volume**	**0.587**	**>0.001**
Left-Amygdala	**0.768**	**>0.001**	**lh_frontalpole_volume**	**0.204**	**0.007**
CSF	**-0.528**	**>0.001**	**lh_temporalpole_volume**	**0.203**	**0.007**
Left-Accumbens-area	**0.458**	**>0.001**	**lh_transversetemporal_volume**	**0.276**	**>0.001**
Left-VentralDC	**0.277**	**>0.001**	**lh_insula_volume**	**0.401**	**>0.001**
Right-Lateral-Ventricle	**-0.533**	**>0.001**	**rh_bankssts_volume**	**0.554**	**>0.001**
Right-Inf-Lat-Vent	**-0.705**	**>0.001**	rh_caudalanteriorcingulate_volume	0.082	0.160
Right-Cerebellum-White-Matter	-0.080	0.166	**rh_caudalmiddlefrontal_volume**	**0.442**	**>0.001**
Right-Cerebellum-Cortex	**0.153**	**0.032**	**rh_cuneus_volume**	**0.148**	**0.037**
Right-Thalamus-Proper	**0.499**	**>0.001**	**rh_entorhinal_volume**	**0.430**	**>0.001**
Right-Caudate	**0.136**	**0.049**	**rh_fusiform_volume**	**0.698**	**>0.001**
Right-Putamen	**0.371**	**>0.001**	**rh_inferiorparietal_volume**	**0.618**	**>0.001**
Right-Pallidum	0.002	0.491	**rh_inferiortemporal_volume**	**0.621**	**>0.001**
Right-Hippocampus	**0.719**	**>0.001**	**rh_isthmuscingulate_volume**	**0.412**	**>0.001**
Right-Amygdala	**0.685**	**>0.001**	**rh_lateraloccipital_volume**	**0.422**	**>0.001**
Right-Accumbens-area	**0.349**	**>0.001**	**rh_lateralorbitofrontal_volume**	**0.422**	**>0.001**
Right-VentralDC	**0.230**	**0.003**	**rh_lingual_volume**	**0.231**	**0.003**
CC_Posterior	-0.026	0.376	**rh_medialorbitofrontal_volume**	**0.415**	**>0.001**
CC_Mid_Posterior	**0.177**	**0.016**	**rh_middletemporal_volume**	**0.697**	**>0.001**
CC_Central	**0.383**	**>0.001**	**rh_parahippocampal_volume**	**0.416**	**>0.001**
CC_Mid_Anterior	**0.306**	**>0.001**	**rh_paracentral_volume**	**0.264**	**0.001**
CC_Anterior	0.059	0.238	**rh_parsopercularis_volume**	**0.310**	**>0.001**
lh_bankssts_volume	**0.656**	**>0.001**	**rh_parsorbitalis_volume**	**0.349**	**>0.001**
lh_caudalanteriorcingulate_volume	-0.035	0.337	**rh_parstriangularis_volume**	**0.266**	**0.001**
lh_caudalmiddlefrontal_volume	**0.487**	**>0.001**	**rh_pericalcarine_volume**	**0.211**	**0.005**
lh_cuneus_volume	**0.200**	**0.008**	**rh_postcentral_volume**	**0.318**	**>0.001**
lh_entorhinal_volume	**0.514**	**>0.001**	**rh_posteriorcingulate_volume**	**0.427**	**>0.001**
lh_fusiform_volume	**0.621**	**>0.001**	**rh_precentral_volume**	**0.280**	**>0.001**
lh_inferiorparietal_volume	**0.644**	**>0.001**	**rh_precuneus_volume**	**0.573**	**>0.001**
lh_inferiortemporal_volume	**0.641**	**>0.001**	**rh_rostralanteriorcingulate_volume**	**0.143**	**0.041**
lh_isthmuscingulate_volume	**0.446**	**>0.001**	**rh_rostralmiddlefrontal_volume**	**0.555**	**>0.001**
lh_lateraloccipital_volume	**0.388**	**>0.001**	**rh_superiorfrontal_volume**	**0.450**	**>0.001**
lh_lateralorbitofrontal_volume	**0.439**	**>0.001**	**rh_superiorparietal_volume**	**0.519**	**>0.001**
lh_lingual_volume	**0.354**	**>0.001**	**rh_superiortemporal_volume**	**0.613**	**>0.001**
lh_medialorbitofrontal_volume	**0.453**	**>0.001**	**rh_supramarginal_volume**	**0.572**	**>0.001**
lh_middletemporal_volume	**0.686**	**>0.001**	rh_frontalpole_volume	0.072	0.190
lh_parahippocampal_volume	**0.458**	**>0.001**	**rh_temporalpole_volume**	**0.297**	**>0.001**
lh_paracentral_volume	**0.250**	**0.001**	**rh_transversetemporal_volume**	**0.230**	**0.003**
lh_parsopercularis_volume	**0.382**	**>0.001**	**rh_insula_volume**	**0.468**	**>0.001**
lh_parsorbitalis_volume	**0.271**	**0.001**			

5. Univariate analysis – AIBL dataset – HC vs MCI

Supplementary Table 4 - Statistical significance measured by the Mann-Whitney U test and effect size measured by Cliff's delta absolute value based on the comparison of the reconstruction error for each brain region between the HC and the MCI groups from the AIBL dataset. The regions with p-value <= 0.05 are highlighted in bold.

Regions	Effect size	p-value	Regions	Effect size	p-value
Left-Lateral-Ventricle	-0.125	0.088	lh_parstriangularis_volume	-0.008	0.466
Left-Inf-Lat-Vent	**-0.273**	**0.002**	lh_pericalcarine_volume	0.109	0.120
Left-Cerebellum-White-Matter	0.128	0.084	lh_postcentral_volume	0.144	0.059
Left-Cerebellum-Cortex	0.140	0.065	lh_posteriorcingulate_volume	0.075	0.210
Left-Thalamus-Proper	0.104	0.132	lh_precentral_volume	0.067	0.236
Left-Caudate	0.115	0.106	**lh_precuneus_volume**	**0.294**	**0.001**
Left-Putamen	-0.013	0.444	lh_rostralanteriorcingulate_volume	0.090	0.164
Left-Pallidum	-0.045	0.313	**lh_rostralmiddlefrontal_volume**	**0.205**	**0.013**
3rd-Ventricle	**-0.234**	**0.006**	**lh_superiorfrontal_volume**	**0.236**	**0.005**
4th-Ventricle	-0.037	0.343	lh_superiorparietal_volume	0.077	0.202
Brain-Stem	0.005	0.480	**lh_superiortemporal_volume**	**0.261**	**0.002**
Left-Hippocampus	**0.299**	**0.001**	lh_supramarginal_volume	0.105	0.129
Left-Amygdala	**0.218**	**0.009**	lh_frontalpole_volume	0.116	0.106
CSF	**-0.323**	**0.000**	**lh_temporalpole_volume**	**0.258**	**0.003**
Left-Accumbens-area	-0.019	0.417	lh_transversetemporal_volume	0.038	0.341
Left-VentralDC	-0.017	0.427	**lh_insula_volume**	**0.223**	**0.008**
Right-Lateral-Ventricle	**-0.167**	**0.035**	rh_bankssts_volume	0.108	0.123
Right-Inf-Lat-Vent	**-0.364**	**>0.001**	rh_caudalanteriorcingulate_volume	0.000	0.500
Right-Cerebellum-White-Matter	0.145	0.058	rh_caudalmiddlefrontal_volume	0.150	0.053
Right-Cerebellum-Cortex	**0.159**	**0.042**	rh_cuneus_volume	0.108	0.122
Right-Thalamus-Proper	**0.165**	**0.037**	rh_entorhinal_volume	0.150	0.052
Right-Caudate	0.029	0.379	**rh_fusiform_volume**	**0.243**	**0.004**
Right-Putamen	-0.001	0.496	**rh_inferiorparietal_volume**	**0.173**	**0.031**
Right-Pallidum	0.020	0.413	rh_inferiortemporal_volume	0.140	0.065
Right-Hippocampus	**0.311**	**>0.001**	**rh_isthmuscingulate_volume**	**0.260**	**0.003**
Right-Amygdala	**0.210**	**0.012**	rh_lateraloccipital_volume	0.129	0.082
Right-Accumbens-area	-0.025	0.392	**rh_lateralorbitofrontal_volume**	**0.253**	**0.003**
Right-VentralDC	-0.077	0.201	**rh_lingual_volume**	**0.189**	**0.020**
CC_Posterior	0.008	0.467	rh_medialorbitofrontal_volume	0.115	0.106
CC_Mid_Posterior	0.117	0.104	**rh_middletemporal_volume**	**0.176**	**0.029**
CC_Central	0.048	0.302	**rh_parahippocampal_volume**	**0.154**	**0.048**
CC_Mid_Anterior	0.118	0.101	rh_paracentral_volume	0.134	0.073
CC_Anterior	0.062	0.250	rh_parsopercularis_volume	0.106	0.126
lh_bankssts_volume	**0.203**	**0.014**	rh_parsorbitalis_volume	0.055	0.278
lh_caudalanteriorcingulate_volume	0.030	0.374	rh_parstriangularis_volume	0.093	0.158
lh_caudalmiddlefrontal_volume	**0.189**	**0.021**	rh_pericalcarine_volume	0.077	0.203
lh_cuneus_volume	**0.193**	**0.018**	**rh_postcentral_volume**	**0.173**	**0.031**
lh_entorhinal_volume	**0.340**	**>0.001**	rh_posteriorcingulate_volume	0.063	0.247
lh_fusiform_volume	**0.249**	**0.003**	rh_precentral_volume	0.131	0.079
lh_inferiorparietal_volume	0.089	0.168	**rh_precuneus_volume**	**0.223**	**0.008**
lh_inferiortemporal_volume	0.143	0.061	**rh_rostralanteriorcingulate_volume**	**0.177**	**0.028**
lh_isthmuscingulate_volume	0.136	0.071	rh_rostralmiddlefrontal_volume	0.091	0.162
lh_lateraloccipital_volume	0.142	0.062	**rh_superiorfrontal_volume**	**0.250**	**0.003**
lh_lateralorbitofrontal_volume	**0.248**	**0.004**	**rh_superiorparietal_volume**	**0.241**	**0.005**
lh_lingual_volume	0.126	0.087	**rh_superiortemporal_volume**	**0.239**	**0.005**
lh_medialorbitofrontal_volume	**0.292**	**0.001**	rh_supramarginal_volume	0.092	0.160
lh_middletemporal_volume	0.120	0.097	rh_frontalpole_volume	0.081	0.189
lh_parahippocampal_volume	0.095	0.151	**rh_temporalpole_volume**	**0.188**	**0.021**
lh_paracentral_volume	0.101	0.138	**rh_transversetemporal_volume**	**0.197**	**0.017**
lh_parsopercularis_volume	0.049	0.298	**rh_insula_volume**	**0.252**	**0.003**
lh_parsorbitalis_volume	0.038	0.341			

6. Univariate analysis – AIBL dataset – HC vs AD

Supplementary Table 5 - Statistical significance measured by the Mann-Whitney U test and effect size measured by Cliff's delta absolute value based on the comparison of the reconstruction error for each brain region between the HC and the AD groups from AIBL dataset. The regions with p-value <= 0.05 are highlighted in bold.

Regions	Effect size	p-value	Regions	Effect size	p-value
Left-Lateral-Ventricle	**-0.584**	**>0.001**	**lh_parstriangularis_volume**	**0.275**	**0.003**
Left-Inf-Lat-Vent	**-0.803**	**>0.001**	lh_pericalcarine_volume	0.133	0.092
Left-Cerebellum-White-Matter	**0.224**	**0.013**	lh_postcentral_volume	0.137	0.086
Left-Cerebellum-Cortex	**0.307**	**0.001**	**lh_posteriorcingulate_volume**	**0.416**	**0.000**
Left-Thalamus-Proper	**0.335**	**>0.001**	lh_precentral_volume	0.075	0.228
Left-Caudate	0.147	0.072	**lh_precuneus_volume**	**0.721**	**>0.001**
Left-Putamen	**0.284**	**0.002**	**lh_rostralanteriorcingulate_volume**	**0.387**	**>0.001**
Left-Pallidum	0.151	0.066	**lh_rostralmiddlefrontal_volume**	**0.510**	**>0.001**
3rd-Ventricle	**-0.510**	**>0.001**	**lh_superiorfrontal_volume**	**0.326**	**0.001**
4th-Ventricle	-0.079	0.216	**lh_superiorparietal_volume**	**0.447**	**>0.001**
Brain-Stem	**0.197**	**0.025**	**lh_superiortemporal_volume**	**0.481**	**>0.001**
Left-Hippocampus	**0.729**	**>0.001**	**lh_supramarginal_volume**	**0.382**	**>0.001**
Left-Amygdala	**0.666**	**>0.001**	**lh_frontalpole_volume**	**0.184**	**0.034**
CSF	**-0.492**	**>0.001**	**lh_temporalpole_volume**	**0.238**	**0.009**
Left-Accumbens-area	**0.217**	**0.015**	lh_transversetemporal_volume	0.113	0.130
Left-VentralDC	**0.226**	**0.012**	**lh_insula_volume**	**0.272**	**0.003**
Right-Lateral-Ventricle	**-0.610**	**>0.001**	**rh_bankssts_volume**	**0.540**	**0.000**
Right-Inf-Lat-Vent	**-0.764**	**>0.001**	**rh_caudalanteriorcingulate_volume**	**0.241**	**0.008**
Right-Cerebellum-White-Matter	**0.203**	**0.022**	**rh_caudalmiddlefrontal_volume**	**0.386**	**>0.001**
Right-Cerebellum-Cortex	**0.313**	**0.001**	**rh_cuneus_volume**	**0.287**	**0.002**
Right-Thalamus-Proper	**0.356**	**>0.001**	**rh_entorhinal_volume**	**0.734**	**>0.001**
Right-Caudate	0.151	0.067	**rh_fusiform_volume**	**0.533**	**>0.001**
Right-Putamen	**0.276**	**0.003**	**rh_inferiorparietal_volume**	**0.728**	**>0.001**
Right-Pallidum	0.141	0.080	**rh_inferiortemporal_volume**	**0.720**	**>0.001**
Right-Hippocampus	**0.751**	**>0.001**	**rh_isthmuscingulate_volume**	**0.595**	**>0.001**
Right-Amygdala	**0.737**	**>0.001**	**rh_lateraloccipital_volume**	**0.485**	**>0.001**
Right-Accumbens-area	**0.362**	**>0.001**	**rh_lateralorbitofrontal_volume**	**0.407**	**>0.001**
Right-VentralDC	**0.270**	**0.004**	**rh_lingual_volume**	**0.339**	**>0.001**
CC_Posterior	0.070	0.244	**rh_medialorbitofrontal_volume**	**0.340**	**>0.001**
CC_Mid_Posterior	**0.307**	**0.001**	**rh_middletemporal_volume**	**0.609**	**>0.001**
CC_Central	**0.393**	**>0.001**	**rh_parahippocampal_volume**	**0.576**	**>0.001**
CC_Mid_Anterior	**0.427**	**>0.001**	rh_paracentral_volume	0.103	0.153
CC_Anterior	0.084	0.203	**rh_parsopercularis_volume**	**0.305**	**0.001**
lh_bankssts_volume	**0.389**	**>0.001**	**rh_parsorbitalis_volume**	**0.326**	**0.001**
lh_caudalanteriorcingulate_volume	0.045	0.326	**rh_parstriangularis_volume**	**0.309**	**0.001**
lh_caudalmiddlefrontal_volume	**0.312**	**0.001**	rh_pericalcarine_volume	0.161	0.054
lh_cuneus_volume	**0.357**	**>0.001**	rh_postcentral_volume	0.046	0.322
lh_entorhinal_volume	**0.719**	**>0.001**	**rh_posteriorcingulate_volume**	**0.360**	**>0.001**
lh_fusiform_volume	**0.586**	**>0.001**	rh_precentral_volume	0.074	0.231
lh_inferiorparietal_volume	**0.636**	**>0.001**	**rh_precuneus_volume**	**0.652**	**>0.001**
lh_inferiortemporal_volume	**0.585**	**>0.001**	**rh_rostralanteriorcingulate_volume**	**0.388**	**>0.001**
lh_isthmuscingulate_volume	**0.590**	**>0.001**	**rh_rostralmiddlefrontal_volume**	**0.469**	**>0.001**
lh_lateraloccipital_volume	**0.589**	**>0.001**	**rh_superiorfrontal_volume**	**0.304**	**0.001**
lh_lateralorbitofrontal_volume	**0.407**	**>0.001**	**rh_superiorparietal_volume**	**0.596**	**>0.001**
lh_lingual_volume	**0.283**	**0.002**	**rh_superiortemporal_volume**	**0.441**	**>0.001**
lh_medialorbitofrontal_volume	**0.214**	**0.016**	**rh_supramarginal_volume**	**0.496**	**>0.001**
lh_middletemporal_volume	**0.624**	**>0.001**	rh_frontalpole_volume	0.092	0.179
lh_parahippocampal_volume	**0.485**	**>0.001**	**rh_temporalpole_volume**	**0.381**	**>0.001**
lh_paracentral_volume	0.137	0.087	**rh_transversetemporal_volume**	**0.219**	**0.014**
lh_parsopercularis_volume	**0.365**	**>0.001**	**rh_insula_volume**	**0.439**	**>0.001**
lh_parsorbitalis_volume	**0.256**	**0.005**			

7. Univariate analysis – ARWIBO dataset – HC vs MCI

Supplementary Table 6 - Statistical significance measured by the Mann-Whitney U test and effect size measured by Cliff's delta absolute value based on the comparison of the reconstruction error for each brain region between the HC and the MCI groups from the ARWIBO dataset. The regions with p-value <= 0.05 are highlighted in bold.

Regions	Effect size	p-value	Regions	Effect size	p-value
Left-Lateral-Ventricle	**-0.251**	**0.001**	**lh_parstriangularis_volume**	**0.199**	**0.007**
Left-Inf-Lat-Vent	**-0.250**	**0.001**	lh_pericalcarine_volume	-0.018	0.412
Left-Cerebellum-White-Matter	**0.167**	**0.020**	lh_postcentral_volume	0.043	0.296
Left-Cerebellum-Cortex	-0.085	0.147	**lh_posteriorcingulate_volume**	**0.236**	**0.002**
Left-Thalamus-Proper	**0.210**	**0.005**	**lh_precentral_volume**	**0.145**	**0.036**
Left-Caudate	0.047	0.280	**lh_precuneus_volume**	**0.229**	**0.002**
Left-Putamen	0.084	0.149	**lh_rostralanteriorcingulate_volume**	**0.139**	**0.043**
Left-Pallidum	0.010	0.452	**lh_rostralmiddlefrontal_volume**	**0.189**	**0.010**
3rd-Ventricle	**-0.235**	**0.002**	**lh_superiorfrontal_volume**	**0.226**	**0.003**
4th-Ventricle	**-0.170**	**0.018**	lh_superiorparietal_volume	0.132	0.051
Brain-Stem	**0.148**	**0.034**	lh_superiortemporal_volume	-0.020	0.402
Left-Hippocampus	**0.276**	**>0.001**	lh_supramarginal_volume	0.070	0.193
Left-Amygdala	0.122	0.066	lh_frontalpole_volume	-0.037	0.323
CSF	**-0.285**	**>0.001**	lh_temporalpole_volume	0.071	0.190
Left-Accumbens-area	**0.277**	**>0.001**	lh_transversetemporal_volume	-0.132	0.051
Left-VentralDC	0.119	0.071	**lh_insula_volume**	**0.185**	**0.011**
Right-Lateral-Ventricle	**-0.261**	**0.001**	**rh_bankssts_volume**	**0.247**	**0.001**
Right-Inf-Lat-Vent	**-0.335**	**>0.001**	rh_caudalanteriorcingulate_volume	0.039	0.315
Right-Cerebellum-White-Matter	0.037	0.325	**rh_caudalmiddlefrontal_volume**	**0.243**	**0.001**
Right-Cerebellum-Cortex	-0.048	0.276	rh_cuneus_volume	-0.026	0.376
Right-Thalamus-Proper	**0.170**	**0.018**	**rh_entorhinal_volume**	**0.272**	**>0.001**
Right-Caudate	0.110	0.087	**rh_fusiform_volume**	**0.306**	**>0.001**
Right-Putamen	0.095	0.121	**rh_inferiorparietal_volume**	**0.271**	**>0.001**
Right-Pallidum	0.085	0.148	**rh_inferiortemporal_volume**	**0.344**	**>0.001**
Right-Hippocampus	**0.269**	**>0.001**	**rh_isthmuscingulate_volume**	**0.137**	**0.045**
Right-Amygdala	**0.237**	**0.002**	**rh_lateraloccipital_volume**	**0.135**	**0.048**
Right-Accumbens-area	0.109	0.088	**rh_lateralorbitofrontal_volume**	**0.195**	**0.008**
Right-VentralDC	0.106	0.095	**rh_lingual_volume**	**0.165**	**0.021**
CC_Posterior	0.110	0.088	**rh_medialorbitofrontal_volume**	**0.232**	**0.002**
CC_Mid_Posterior	0.120	0.069	**rh_middletemporal_volume**	**0.228**	**0.002**
CC_Central	**0.147**	**0.035**	**rh_parahippocampal_volume**	**0.228**	**0.002**
CC_Mid_Anterior	0.089	0.135	**rh_paracentral_volume**	**0.166**	**0.020**
CC_Anterior	0.068	0.202	rh_parsopercularis_volume	0.098	0.114
lh_bankssts_volume	0.085	0.146	**rh_parsorbitalis_volume**	**0.181**	**0.013**
lh_caudalanteriorcingulate_volume	0.125	0.062	**rh_parstriangularis_volume**	**0.168**	**0.019**
lh_caudalmiddlefrontal_volume	**0.147**	**0.035**	rh_pericalcarine_volume	0.093	0.126
lh_cuneus_volume	-0.025	0.378	rh_postcentral_volume	-0.020	0.400
lh_entorhinal_volume	0.203	0.006	**rh_posteriorcingulate_volume**	**0.244**	**0.001**
lh_fusiform_volume	**0.262**	**0.001**	**rh_precentral_volume**	**0.198**	**0.007**
lh_inferiorparietal_volume	0.198	0.007	**rh_precuneus_volume**	**0.256**	**0.001**
lh_inferiortemporal_volume	**0.275**	**>0.001**	**rh_rostralanteriorcingulate_volume**	**0.191**	**0.009**
lh_isthmuscingulate_volume	**0.155**	**0.028**	rh_rostralmiddlefrontal_volume	0.119	0.070
lh_lateraloccipital_volume	**0.176**	**0.015**	**rh_superiorfrontal_volume**	**0.261**	**0.001**
lh_lateralorbitofrontal_volume	**0.193**	**0.009**	**rh_superiorparietal_volume**	**0.204**	**0.006**
lh_lingual_volume	**0.175**	**0.015**	rh_superiortemporal_volume	0.097	0.117
lh_medialorbitofrontal_volume	**0.253**	**0.001**	rh_supramarginal_volume	0.131	0.053
lh_middletemporal_volume	**0.280**	**>0.001**	rh_frontalpole_volume	-0.023	0.389
lh_parahippocampal_volume	**0.325**	**>0.001**	rh_temporalpole_volume	0.084	0.150
lh_paracentral_volume	0.086	0.145	rh_transversetemporal_volume	0.048	0.276
lh_parsopercularis_volume	**0.163**	**0.022**	**rh_insula_volume**	**0.149**	**0.033**
lh_parsorbitalis_volume	**0.209**	**0.005**			

8. Univariate analysis – ARWIBO dataset – HC vs AD

Supplementary Table 7 - Statistical significance measured by the Mann-Whitney U test and effect size measured by Cliff's delta absolute value based on the comparison of the reconstruction error for each brain region between the HC and the AD groups from ARWIBO dataset. The regions with p-value <= 0.05 are highlighted in bold.

Regions	Effect size	p-value	Regions	Effect size	p-value
Left-Lateral-Ventricle	**-0.547**	**>0.001**	**lh_parstriangularis_volume**	**0.233**	**0.012**
Left-Inf-Lat-Vent	**-0.758**	**>0.001**	lh_pericalcarine_volume	0.080	0.220
Left-Cerebellum-White-Matter	0.063	0.272	lh_postcentral_volume	-0.123	0.117
Left-Cerebellum-Cortex	-0.017	0.435	**lh_posteriorcingulate_volume**	**0.435**	**>0.001**
Left-Thalamus-Proper	**0.282**	**0.003**	lh_precentral_volume	-0.029	0.391
Left-Caudate	**0.321**	**0.001**	**lh_precuneus_volume**	**0.599**	**>0.001**
Left-Putamen	**0.287**	**0.003**	**lh_rostralanteriorcingulate_volume**	**0.289**	**0.003**
Left-Pallidum	-0.156	0.065	**lh_rostralmiddlefrontal_volume**	**0.373**	**>0.001**
3rd-Ventricle	**-0.517**	**>0.001**	**lh_superiorfrontal_volume**	**0.362**	**>0.001**
4th-Ventricle	-0.050	0.314	**lh_superiorparietal_volume**	**0.303**	**0.002**
Brain-Stem	0.126	0.112	**lh_superiortemporal_volume**	**0.382**	**>0.001**
Left-Hippocampus	**0.790**	**>0.001**	**lh_supramarginal_volume**	**0.342**	**>0.001**
Left-Amygdala	**0.649**	**>0.001**	**lh_frontalpole_volume**	**0.211**	**0.020**
CSF	**-0.518**	**>0.001**	**lh_temporalpole_volume**	**0.477**	**>0.001**
Left-Accumbens-area	**0.325**	**0.001**	**lh_transversetemporal_volume**	**0.216**	**0.018**
Left-VentralDC	0.138	0.091	**lh_insula_volume**	**0.510**	**>0.001**
Right-Lateral-Ventricle	**-0.478**	**>0.001**	**rh_bankssts_volume**	**0.486**	**>0.001**
Right-Inf-Lat-Vent	**-0.768**	**>0.001**	rh_caudalanteriorcingulate_volume	0.042	0.342
Right-Cerebellum-White-Matter	0.098	0.170	rh_caudalmiddlefrontal_volume	0.129	0.105
Right-Cerebellum-Cortex	-0.024	0.409	rh_cuneus_volume	-0.009	0.467
Right-Thalamus-Proper	0.117	0.129	**rh_entorhinal_volume**	**0.603**	**>0.001**
Right-Caudate	**0.301**	**0.002**	**rh_fusiform_volume**	**0.506**	**>0.001**
Right-Putamen	**0.272**	**0.004**	**rh_inferiorparietal_volume**	**0.599**	**>0.001**
Right-Pallidum	-0.042	0.344	**rh_inferiortemporal_volume**	**0.415**	**>0.001**
Right-Hippocampus	**0.635**	**>0.001**	**rh_isthmuscingulate_volume**	**0.370**	**>0.001**
Right-Amygdala	**0.651**	**>0.001**	**rh_lateraloccipital_volume**	**0.397**	**>0.001**
Right-Accumbens-area	**0.309**	**0.001**	**rh_lateralorbitofrontal_volume**	**0.305**	**0.002**
Right-VentralDC	0.093	0.184	**rh_lingual_volume**	**0.331**	**0.001**
CC_Posterior	0.155	0.066	**rh_medialorbitofrontal_volume**	**0.303**	**0.002**
CC_Mid_Posterior	0.008	0.469	**rh_middletemporal_volume**	**0.519**	**>0.001**
CC_Central	**0.258**	**0.006**	**rh_parahippocampal_volume**	**0.326**	**0.001**
CC_Mid_Anterior	**0.289**	**0.002**	rh_paracentral_volume	0.114	0.135
CC_Anterior	0.041	0.345	rh_parsopercularis_volume	0.088	0.197
lh_bankssts_volume	**0.394**	**0.000**	**rh_parsorbitalis_volume**	**0.325**	**0.001**
lh_caudalanteriorcingulate_volume	0.037	0.361	**rh_parstriangularis_volume**	**0.201**	**0.026**
lh_caudalmiddlefrontal_volume	**0.352**	**>0.001**	rh_pericalcarine_volume	0.061	0.276
lh_cuneus_volume	0.037	0.361	rh_postcentral_volume	-0.147	0.077
lh_entorhinal_volume	**0.633**	**>0.001**	**rh_posteriorcingulate_volume**	**0.340**	**>0.001**
lh_fusiform_volume	**0.657**	**>0.001**	rh_precentral_volume	-0.022	0.417
lh_inferiorparietal_volume	**0.526**	**>0.001**	**rh_precuneus_volume**	**0.570**	**>0.001**
lh_inferiortemporal_volume	**0.553**	**>0.001**	**rh_rostralanteriorcingulate_volume**	**0.240**	**0.010**
lh_isthmuscingulate_volume	**0.344**	**>0.001**	**rh_rostralmiddlefrontal_volume**	**0.304**	**0.002**
lh_lateraloccipital_volume	**0.322**	**0.001**	**rh_superiorfrontal_volume**	**0.330**	**0.001**
lh_lateralorbitofrontal_volume	**0.325**	**0.001**	**rh_superiorparietal_volume**	**0.410**	**>0.001**
lh_lingual_volume	**0.302**	**0.002**	**rh_superiortemporal_volume**	**0.350**	**>0.001**
lh_medialorbitofrontal_volume	**0.205**	**0.024**	**rh_supramarginal_volume**	**0.380**	**>0.001**
lh_middletemporal_volume	**0.543**	**>0.001**	rh_frontalpole_volume	0.136	0.094
lh_parahippocampal_volume	**0.501**	**>0.001**	**rh_temporalpole_volume**	**0.365**	**>0.001**
lh_paracentral_volume	-0.002	0.493	rh_transversetemporal_volume	0.141	0.085
lh_parsopercularis_volume	**0.259**	**0.006**	**rh_insula_volume**	**0.426**	**>0.001**
lh_parsorbitalis_volume	**0.422**	**>0.001**			

9. Univariate analysis – OASIS-1 dataset – HC vs AD

Supplementary Table 8 - Statistical significance measured by the Mann-Whitney U test and effect size measured by Cliff's delta absolute value based on the comparison of the reconstruction error for each brain region between the HC and the AD groups from the OASIS-1 dataset. The regions with p-value <= 0.05 are highlighted in bold.

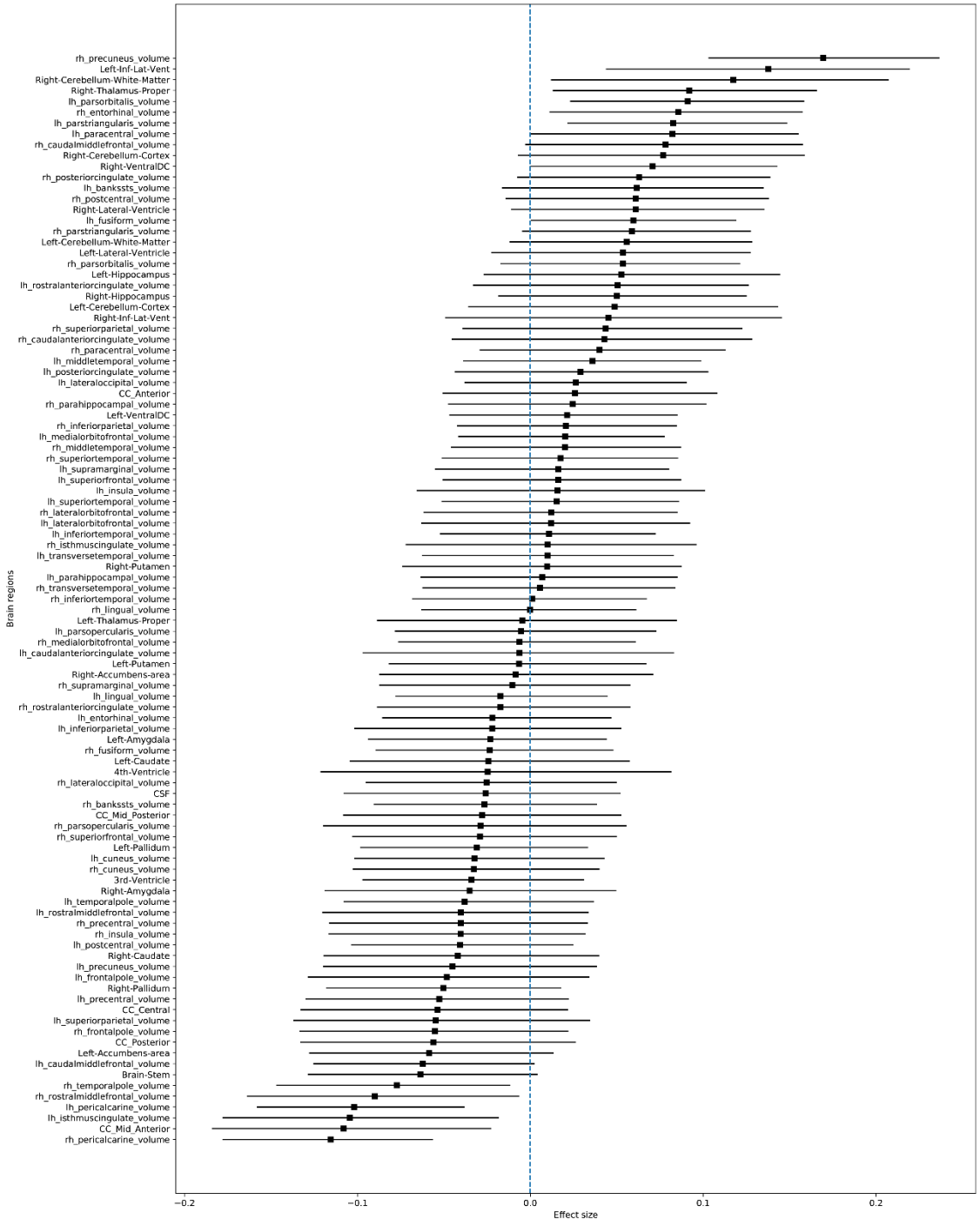
Regions	Effect size	p-value	Regions	Effect size	p-value
Left-Lateral-Ventricle	**-0.469**	**>0.001**	lh_parstriangularis_volume	0.118	0.187
Left-Inf-Lat-Vent	**-0.735**	**>0.001**	lh_pericalcarine_volume	-0.010	0.472
Left-Cerebellum-White-Matter	**0.218**	**0.049**	**lh_postcentral_volume**	**0.362**	**0.003**
Left-Cerebellum-Cortex	**0.287**	**0.015**	**lh_posteriorcingulate_volume**	**0.359**	**0.003**
Left-Thalamus-Proper	**0.492**	**>0.001**	**lh_precentral_volume**	**0.330**	**0.006**
Left-Caudate	0.040	0.382	**lh_precuneus_volume**	**0.434**	**0.000**
Left-Putamen	0.210	0.056	lh_rostralanteriorcingulate_volume	0.102	0.221
Left-Pallidum	0.030	0.413	**lh_rostralmiddlefrontal_volume**	**0.337**	**0.005**
3rd-Ventricle	-0.457	**>0.001**	**lh_superiorfrontal_volume**	**0.358**	**0.003**
4th-Ventricle	0.006	0.484	**lh_superiorparietal_volume**	**0.313**	**0.009**
Brain-Stem	**0.334**	**0.006**	**lh_superiortemporal_volume**	**0.502**	**>0.001**
Left-Hippocampus	**0.494**	**>0.001**	**lh_supramarginal_volume**	**0.502**	**>0.001**
Left-Amygdala	**0.481**	**>0.001**	lh_frontalpole_volume	0.151	0.127
CSF	**-0.350**	**0.004**	lh_temporalpole_volume	0.032	0.405
Left-Accumbens-area	**0.375**	**0.002**	**lh_transversetemporal_volume**	**0.343**	**0.005**
Left-VentralDC	**0.275**	**0.019**	**lh_insula_volume**	**0.313**	**0.009**
Right-Lateral-Ventricle	**-0.516**	**>0.001**	**rh_bankssts_volume**	**0.644**	**>0.001**
Right-Inf-Lat-Vent	**-0.780**	**>0.001**	rh_caudalanteriorcingulate_volume	0.027	0.421
Right-Cerebellum-White-Matter	0.150	0.129	**rh_caudalmiddlefrontal_volume**	**0.387**	**0.002**
Right-Cerebellum-Cortex	**0.227**	**0.043**	rh_cuneus_volume	0.151	0.127
Right-Thalamus-Proper	**0.521**	**>0.001**	**rh_entorhinal_volume**	**0.566**	**>0.001**
Right-Caudate	0.002	0.496	**rh_fusiform_volume**	**0.511**	**>0.001**
Right-Putamen	**0.339**	**0.005**	**rh_inferiorparietal_volume**	**0.462**	**>0.001**
Right-Pallidum	0.007	0.480	**rh_inferiortemporal_volume**	**0.572**	**>0.001**
Right-Hippocampus	**0.623**	**>0.001**	**rh_isthmuscingulate_volume**	**0.274**	**0.019**
Right-Amygdala	**0.541**	**>0.001**	**rh_lateraloccipital_volume**	**0.249**	**0.030**
Right-Accumbens-area	**0.424**	**0.001**	**rh_lateralorbitofrontal_volume**	**0.243**	**0.033**
Right-VentralDC	**0.354**	**0.004**	**rh_lingual_volume**	**0.313**	**0.009**
CC_Posterior	0.183	0.084	**rh_medialorbitofrontal_volume**	**0.312**	**0.009**
CC_Mid_Posterior	**0.496**	**>0.001**	**rh_middletemporal_volume**	**0.499**	**>0.001**
CC_Central	**0.503**	**>0.001**	**rh_parahippocampal_volume**	**0.430**	**0.001**
CC_Mid_Anterior	**0.432**	**0.001**	**rh_paracentral_volume**	**0.234**	**0.038**
CC_Anterior	**0.375**	**0.002**	**rh_parsopercularis_volume**	**0.297**	**0.012**
lh_bankssts_volume	**0.428**	**0.001**	**rh_parsorbitalis_volume**	**0.283**	**0.016**
lh_caudalanteriorcingulate_volume	0.011	0.468	**rh_parstriangularis_volume**	**0.283**	**0.016**
lh_caudalmiddlefrontal_volume	0.113	0.198	rh_pericalcarine_volume	0.074	0.288
lh_cuneus_volume	0.200	0.065	**rh_postcentral_volume**	**0.379**	**0.002**
lh_entorhinal_volume	**0.339**	**0.005**	rh_posteriorcingulate_volume	0.183	0.084
lh_fusiform_volume	**0.375**	**0.002**	**rh_precentral_volume**	**0.428**	**0.001**
lh_inferiorparietal_volume	**0.452**	**>0.001**	**rh_precuneus_volume**	**0.368**	**0.003**
lh_inferiortemporal_volume	**0.456**	**>0.001**	rh_rostralanteriorcingulate_volume	0.114	0.195
lh_isthmuscingulate_volume	**0.444**	**>0.001**	**rh_rostralmiddlefrontal_volume**	**0.425**	**0.001**
lh_lateraloccipital_volume	**0.314**	**0.009**	**rh_superiorfrontal_volume**	**0.425**	**0.001**
lh_lateralorbitofrontal_volume	**0.354**	**0.004**	**rh_superiorparietal_volume**	**0.272**	**0.020**
lh_lingual_volume	**0.238**	**0.036**	**rh_superiortemporal_volume**	**0.548**	**>0.001**
lh_medialorbitofrontal_volume	0.064	0.316	**rh_supramarginal_volume**	**0.519**	**>0.001**
lh_middletemporal_volume	**0.507**	**>0.001**	rh_frontalpole_volume	0.126	0.171
lh_parahippocampal_volume	**0.458**	**>0.001**	rh_temporalpole_volume	0.007	0.480
lh_paracentral_volume	**0.333**	**0.006**	**rh_transversetemporal_volume**	**0.413**	**0.001**
lh_parsopercularis_volume	**0.245**	**0.032**	**rh_insula_volume**	**0.292**	**0.014**
lh_parsorbitalis_volume	**0.270**	**0.021**			

10. Univariate analysis –MIRIAD dataset – HC vs AD

Supplementary Table 9 - Statistical significance measured by the Mann-Whitney U test and effect size measured by Cliff's delta absolute value based on the comparison of the reconstruction error for each brain region between the HC and the AD groups from the MIRIAD dataset. The regions with p-value <= 0.05 are highlighted in bold.

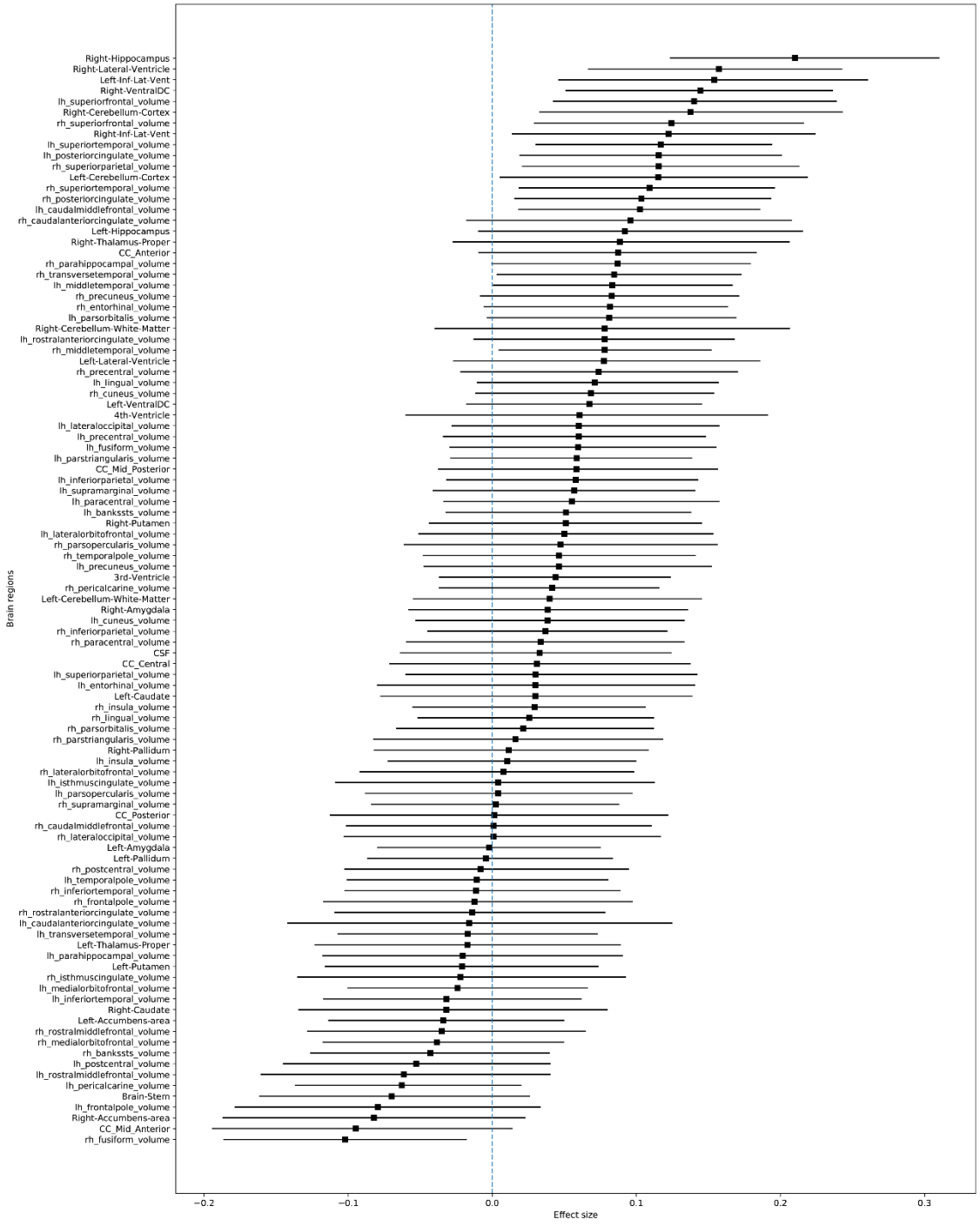
Regions	Effect size	p-value	Regions	Effect size	p-value
Left-Lateral-Ventricle	**-0.630**	**>0.001**	lh_parstriangularis_volume	0.034	0.361
Left-Inf-Lat-Vent	**-0.964**	**>0.001**	lh_pericalcarine_volume	-0.073	0.222
Left-Cerebellum-White-Matter	0.022	0.411	**lh_postcentral_volume**	**0.394**	**>0.001**
Left-Cerebellum-Cortex	**0.396**	**>0.001**	**lh_posteriorcingulate_volume**	**0.545**	**>0.001**
Left-Thalamus-Proper	**0.470**	**>0.001**	**lh_precentral_volume**	**0.430**	**>0.001**
Left-Caudate	**0.338**	**>0.001**	**lh_precuneus_volume**	**0.841**	**>0.001**
Left-Putamen	**0.386**	**>0.001**	**lh_rostralanteriorcingulate_volume**	**0.440**	**>0.001**
Left-Pallidum	0.092	0.169	**lh_rostralmiddlefrontal_volume**	**0.836**	**>0.001**
3rd-Ventricle	**-0.385**	**>0.001**	**lh_superiorfrontal_volume**	**0.856**	**>0.001**
4th-Ventricle	0.059	0.269	**lh_superiorparietal_volume**	**0.741**	**>0.001**
Brain-Stem	**0.334**	**>0.001**	**lh_superiortemporal_volume**	**0.753**	**>0.001**
Left-Hippocampus	**0.780**	**>0.001**	**lh_supramarginal_volume**	**0.788**	**>0.001**
Left-Amygdala	**0.878**	**>0.001**	**lh_frontalpole_volume**	**0.180**	**0.030**
CSF	**-0.673**	**>0.001**	**lh_temporalpole_volume**	**0.367**	**>0.001**
Left-Accumbens-area	**0.526**	**>0.001**	**lh_transversetemporal_volume**	**0.532**	**>0.001**
Left-VentralDC	**0.315**	**>0.001**	**lh_insula_volume**	**0.542**	**>0.001**
Right-Lateral-Ventricle	**-0.723**	**>0.001**	**rh_bankssts_volume**	**0.641**	**>0.001**
Right-Inf-Lat-Vent	**-0.932**	**>0.001**	**rh_caudalanteriorcingulate_volume**	**0.374**	**>0.001**
Right-Cerebellum-White-Matter	0.131	0.086	**rh_caudalmiddlefrontal_volume**	**0.471**	**>0.001**
Right-Cerebellum-Cortex	**0.418**	**>0.001**	**rh_cuneus_volume**	**0.430**	**>0.001**
Right-Thalamus-Proper	**0.416**	**>0.001**	**rh_entorhinal_volume**	**0.593**	**>0.001**
Right-Caudate	**0.183**	**0.028**	**rh_fusiform_volume**	**0.819**	**>0.001**
Right-Putamen	**0.406**	**>0.001**	**rh_inferiorparietal_volume**	**0.807**	**>0.001**
Right-Pallidum	0.060	0.265	**rh_inferiortemporal_volume**	**0.804**	**>0.001**
Right-Hippocampus	**0.802**	**>0.001**	**rh_isthmuscingulate_volume**	**0.712**	**>0.001**
Right-Amygdala	**0.877**	**>0.001**	**rh_lateraloccipital_volume**	**0.457**	**>0.001**
Right-Accumbens-area	**0.558**	**>0.001**	**rh_lateralorbitofrontal_volume**	**0.547**	**>0.001**
Right-VentralDC	**0.422**	**>0.001**	**rh_lingual_volume**	**0.396**	**>0.001**
CC_Posterior	0.110	0.125	**rh_medialorbitofrontal_volume**	**0.481**	**>0.001**
CC_Mid_Posterior	-0.067	0.244	**rh_middletemporal_volume**	**0.825**	**>0.001**
CC_Central	**0.411**	**>0.001**	**rh_parahippocampal_volume**	**0.488**	**>0.001**
CC_Mid_Anterior	**0.500**	**>0.001**	rh_paracentral_volume	0.150	0.058
CC_Anterior	0.083	0.193	**rh_parsopercularis_volume**	**0.363**	**>0.001**
lh_bankssts_volume	**0.791**	**>0.001**	**rh_parsorbitalis_volume**	**0.232**	**0.008**
lh_caudalanteriorcingulate_volume	-0.095	0.162	rh_parstriangularis_volume	-0.110	0.126
lh_caudalmiddlefrontal_volume	**0.643**	**>0.001**	rh_pericalcarine_volume	0.046	0.315
lh_cuneus_volume	0.156	0.052	**rh_postcentral_volume**	**0.238**	**0.006**
lh_entorhinal_volume	**0.446**	**>0.001**	**rh_posteriorcingulate_volume**	**0.575**	**>0.001**
lh_fusiform_volume	**0.818**	**>0.001**	**rh_precentral_volume**	**0.424**	**>0.001**
lh_inferiorparietal_volume	**0.869**	**>0.001**	**rh_precuneus_volume**	**0.913**	**>0.001**
lh_inferiortemporal_volume	**0.884**	**>0.001**	**rh_rostralanteriorcingulate_volume**	**0.245**	**0.005**
lh_isthmuscingulate_volume	**0.604**	**>0.001**	**rh_rostralmiddlefrontal_volume**	**0.804**	**>0.001**
lh_lateraloccipital_volume	**0.628**	**>0.001**	**rh_superiorfrontal_volume**	**0.686**	**>0.001**
lh_lateralorbitofrontal_volume	**0.601**	**>0.001**	**rh_superiorparietal_volume**	**0.812**	**>0.001**
lh_lingual_volume	**0.304**	**0.001**	**rh_superiortemporal_volume**	**0.645**	**>0.001**
lh_medialorbitofrontal_volume	**0.535**	**>0.001**	**rh_supramarginal_volume**	**0.857**	**>0.001**
lh_middletemporal_volume	**0.882**	**>0.001**	**rh_frontalpole_volume**	**0.216**	**0.012**
lh_parahippocampal_volume	0.140	0.072	**rh_temporalpole_volume**	**0.437**	**>0.001**
lh_paracentral_volume	**0.381**	**>0.001**	**rh_transversetemporal_volume**	**0.393**	**>0.001**
lh_parsopercularis_volume	**0.314**	**0.001**	**rh_insula_volume**	**0.717**	**>0.001**
lh_parsorbitalis_volume	**0.333**	**>0.001**			

11. Region importance – ADNI dataset – HC vs EMCI



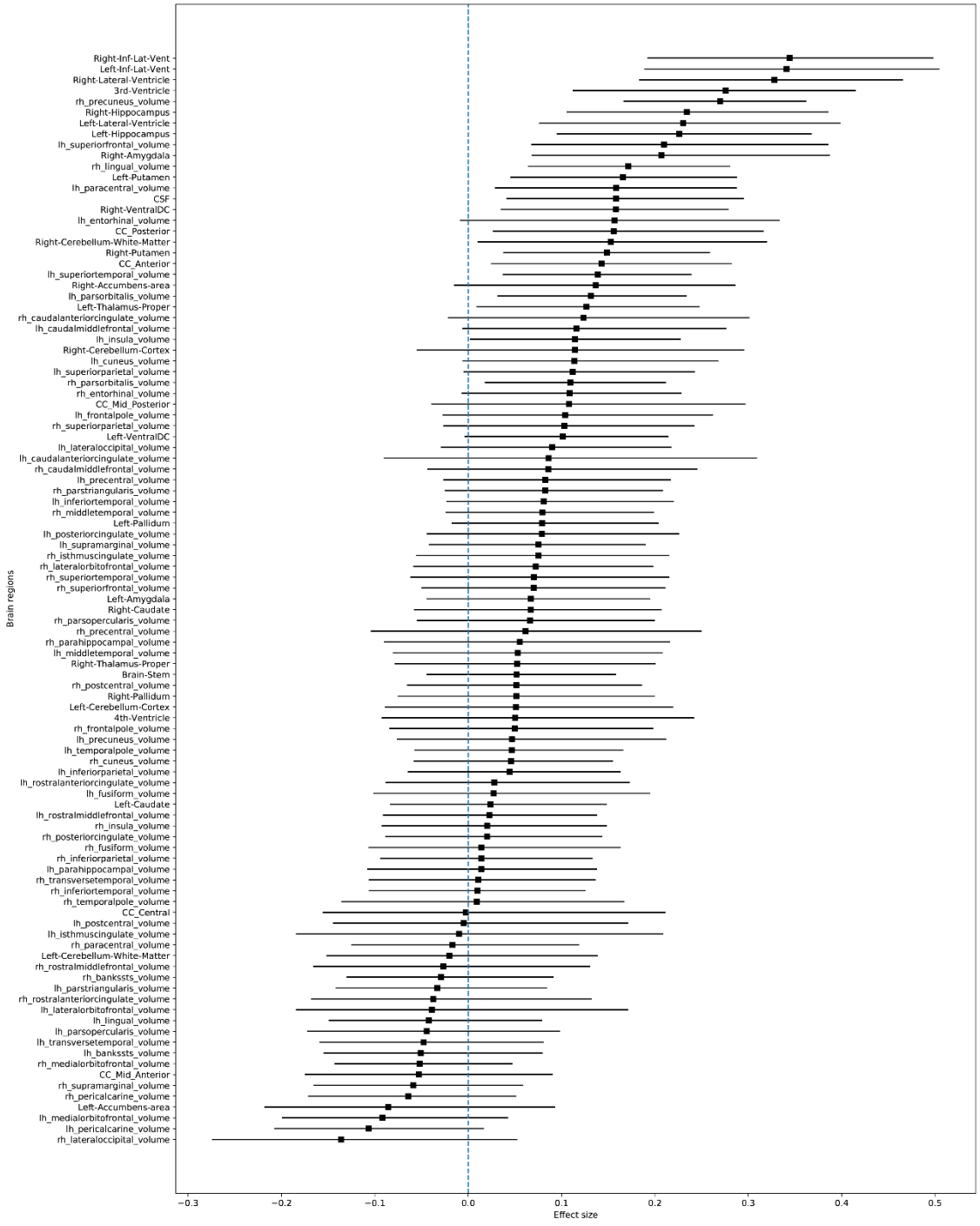
Supplementary Figure 1 - Regional deviations of the EMCI group from the ADNI dataset. The marker indicates the mean effect size between the HC and the EMCI groups. The horizontal bars indicate the 95% confidence interval calculated using the percentile method on the bootstrap analysis.

12. Region importance – ADNI dataset – HC vs LMCI



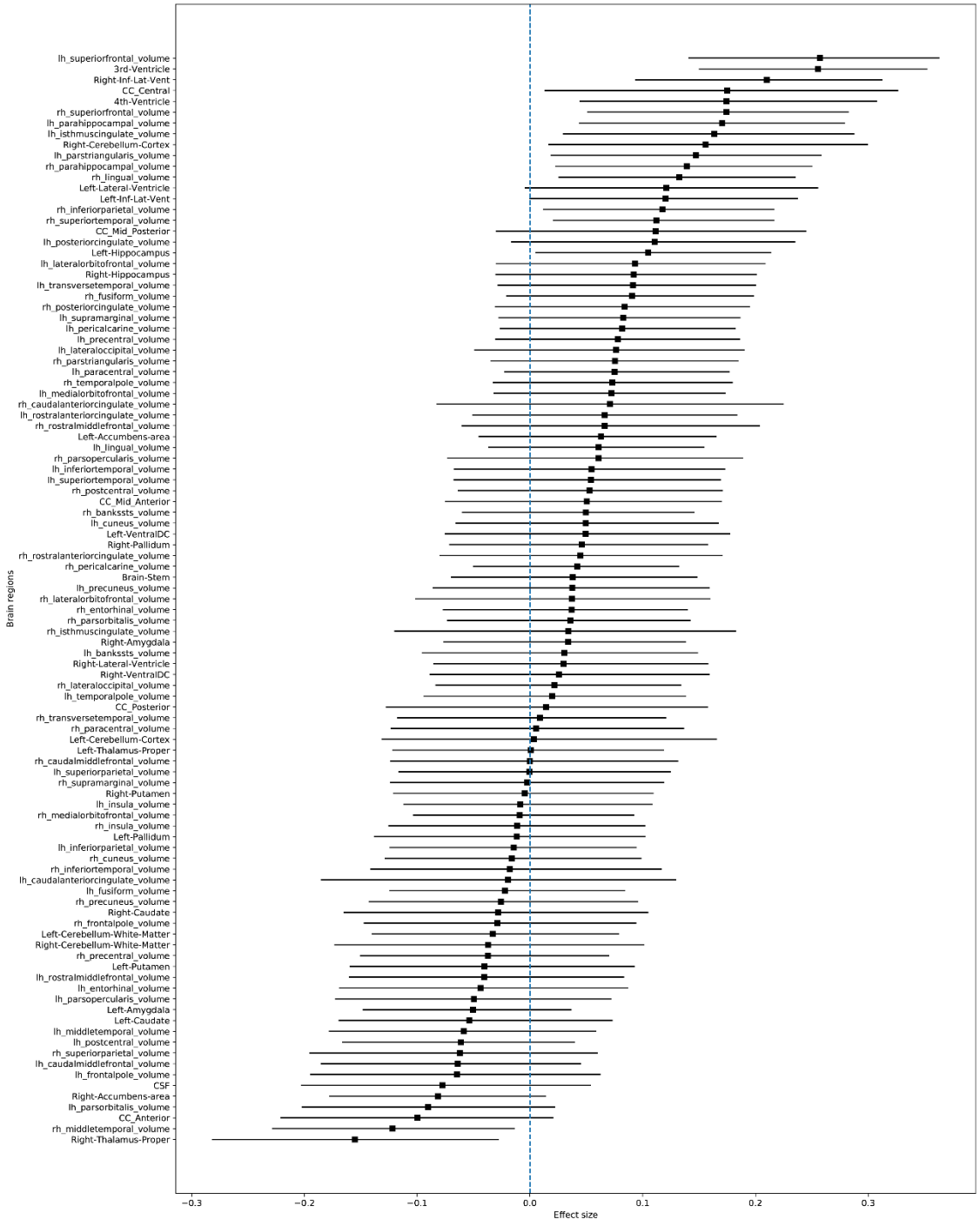
Supplementary Figure 2 – Regional deviations of the LMCI group from the ADNI dataset. The marker indicates the mean effect size between the HC and the LMCI groups. The horizontal bars indicate the 95% confidence interval calculated using the percentile method on the bootstrap analysis.

13. Region importance – ADNI dataset – HC vs AD



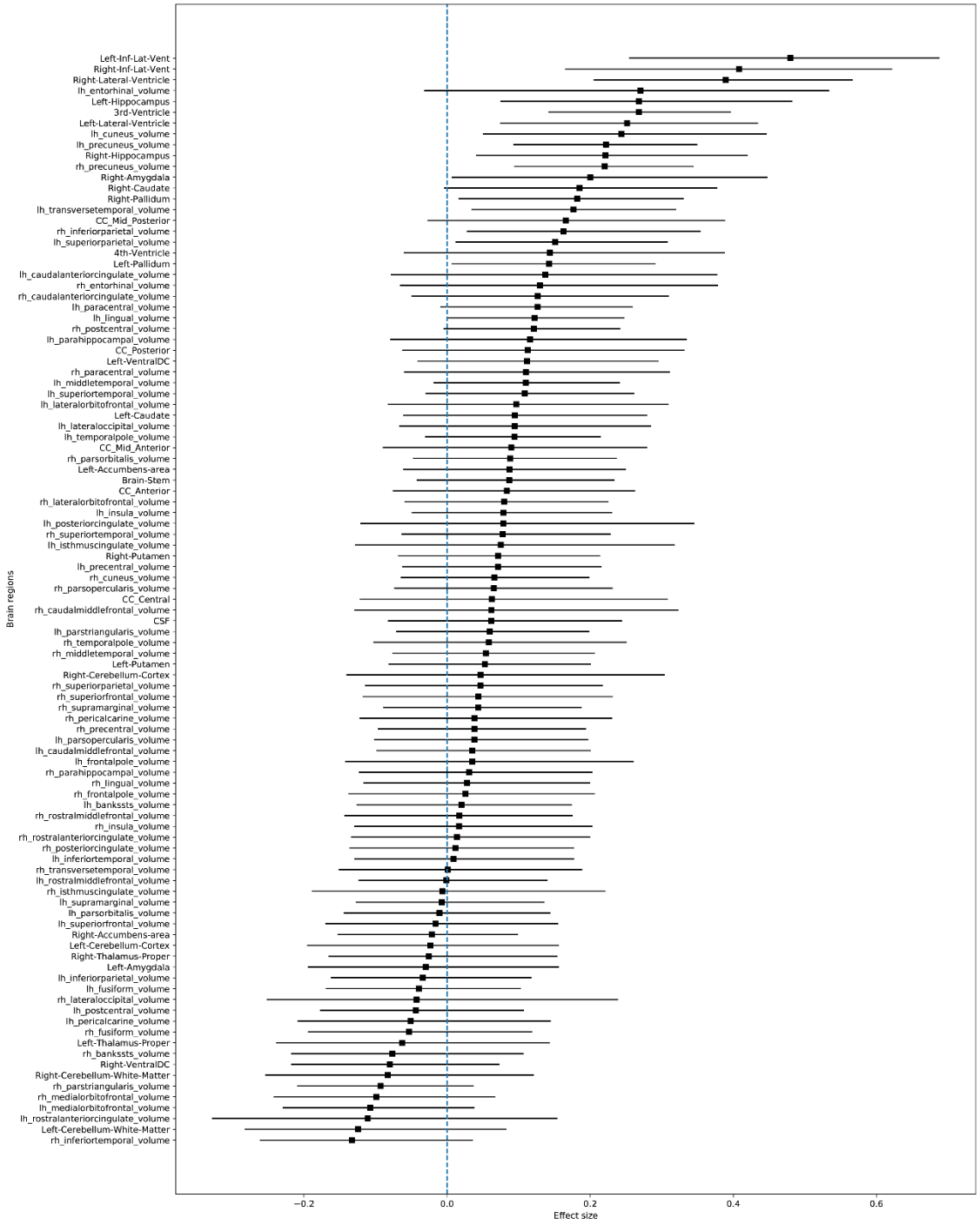
Supplementary Figure 3 – Regional deviations of the AD group from the ADNI dataset. The marker indicates the mean effect size between the HC and the AD groups. The horizontal bars indicate the 95% confidence interval calculated using the percentile method on the bootstrap analysis.

14. Region importance – AIBL dataset – HC vs MCI



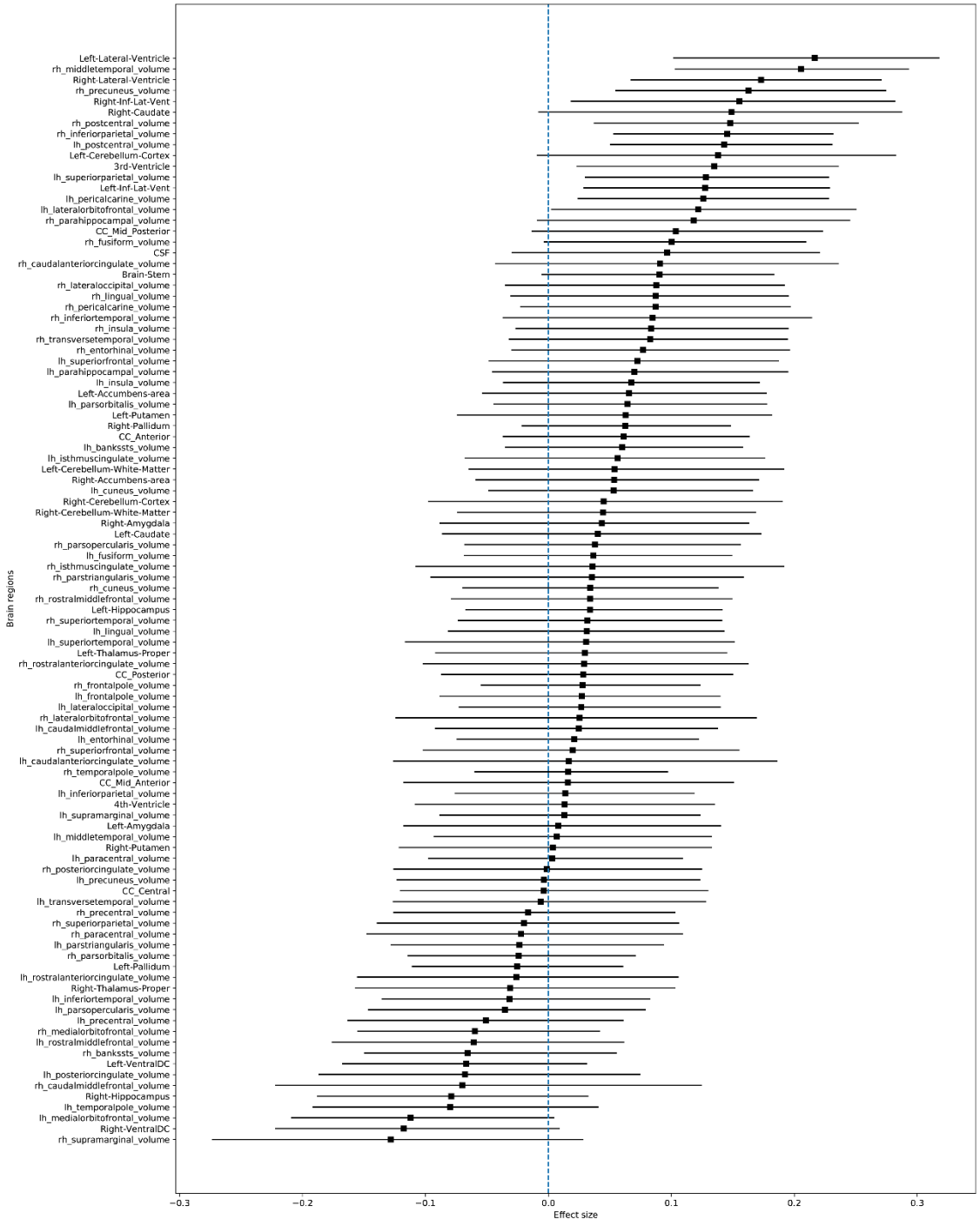
Supplementary Figure 4 – Regional deviations of the MCI group from the AIBL dataset. The marker indicates the mean effect size between the HC and the MCI groups. The horizontal bars indicate the 95% confidence interval calculated using the percentile method on the bootstrap analysis.

15. Region importance – AIBL dataset – HC vs AD



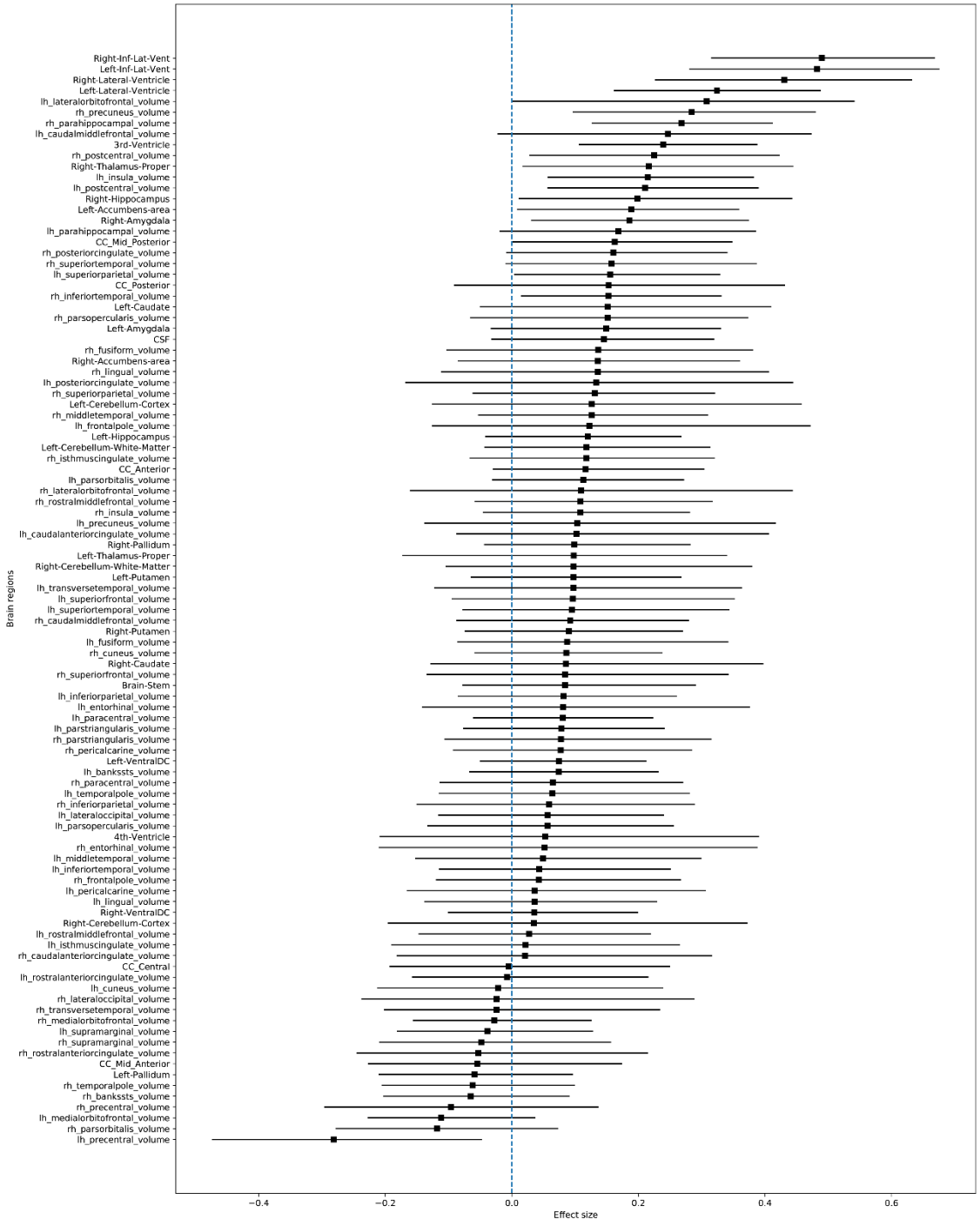
Supplementary Figure 5 – Regional deviations of the AD group from the AIBL dataset. The marker indicates the mean effect size between the HC and the AD groups. The horizontal bars indicate the 95% confidence interval calculated using the percentile method on the bootstrap analysis.

16. Region importance – ARWIBO dataset – HC vs MCI



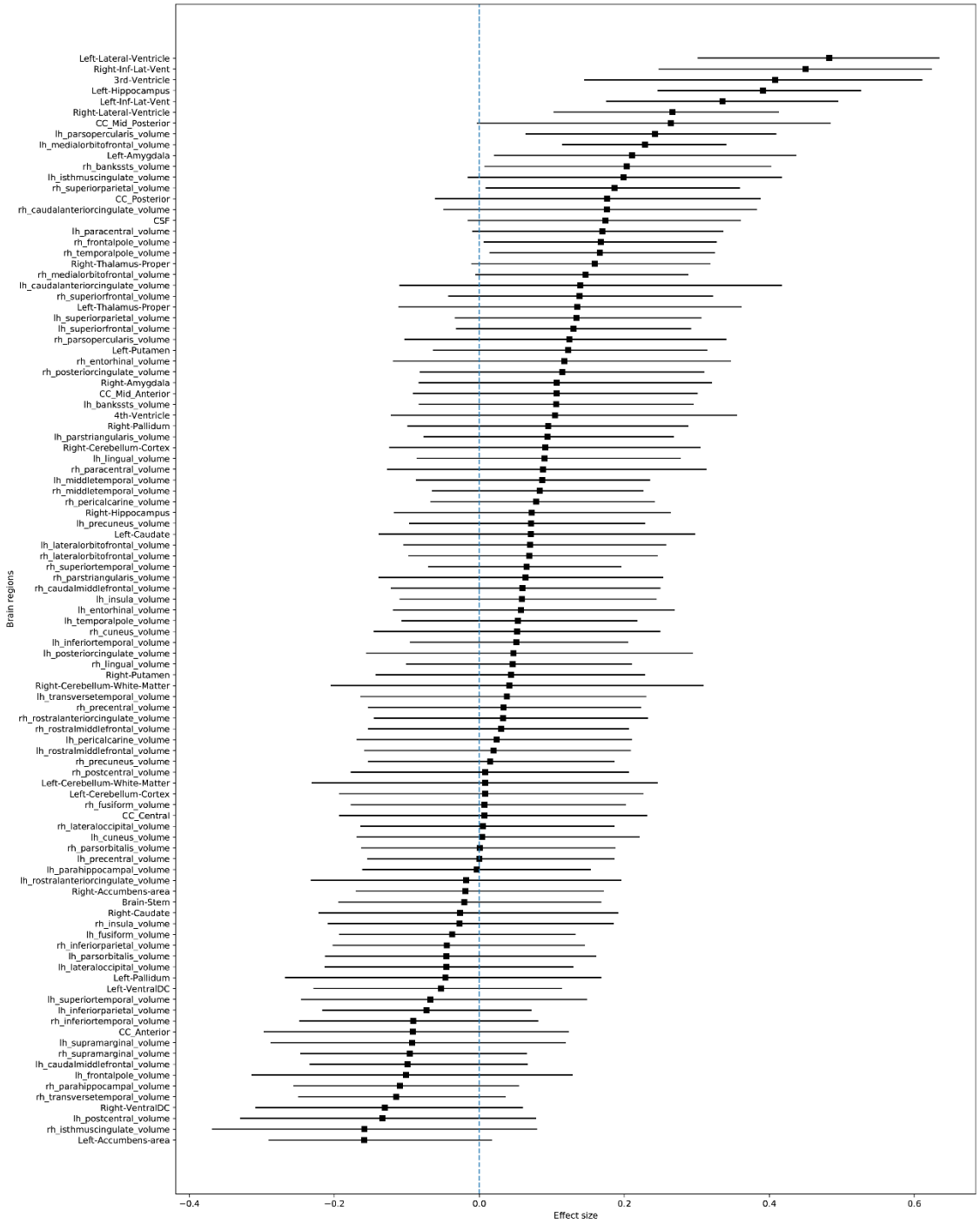
Supplementary Figure 6 - Regional deviations of the MCI group from the ARWIBO dataset. The marker indicates the mean effect size between the HC and the MCI groups. The horizontal bars indicate the 95% confidence interval calculated using the percentile method on the bootstrap analysis.

17. Region importance – ARWIBO dataset – HC vs AD



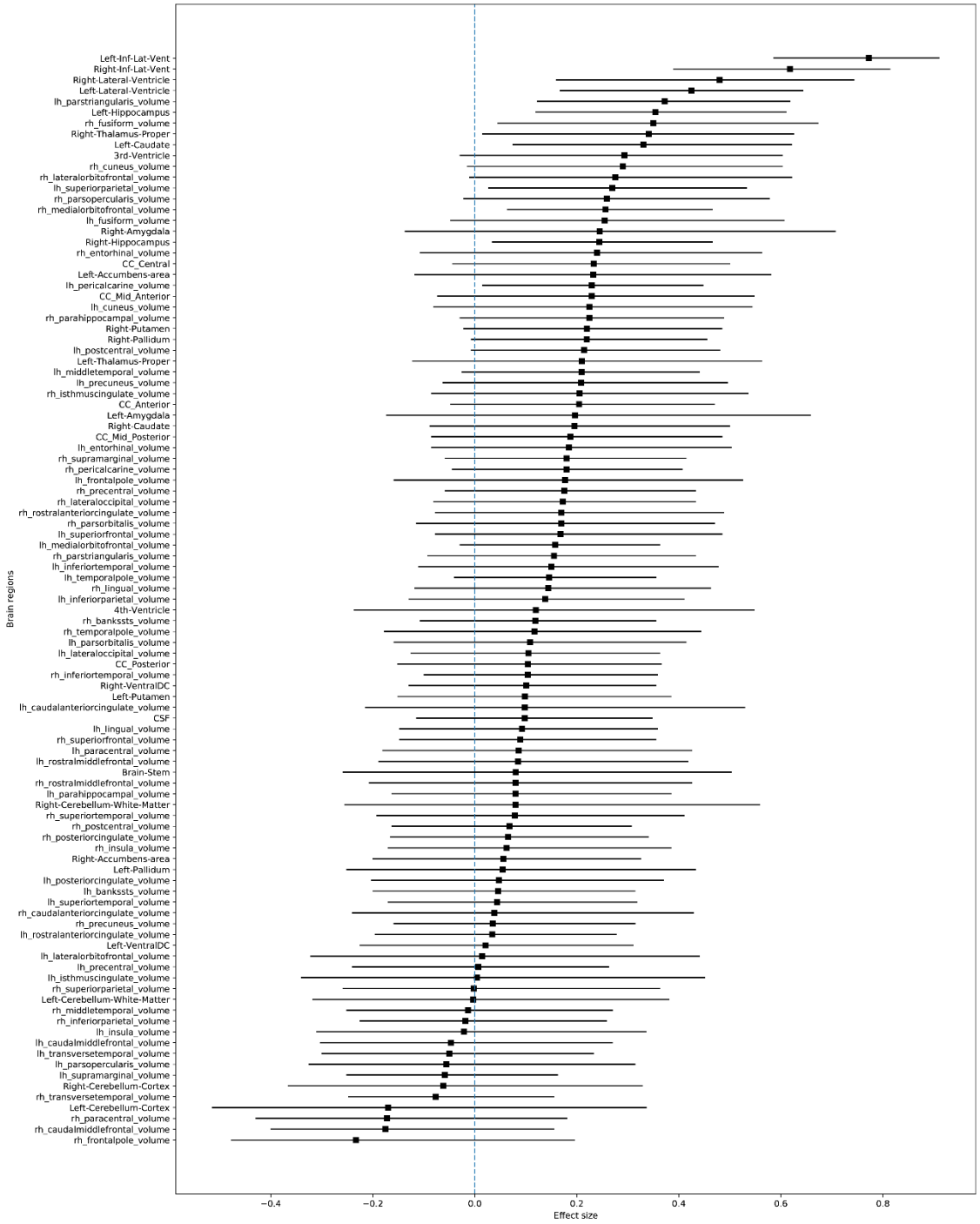
Supplementary Figure 7 – Regional deviations of the AD group from the ARWIBO dataset. The marker indicates the mean effect size between the HC and the AD groups. The horizontal bars indicate the 95% confidence interval calculated using the percentile method on the bootstrap analysis.

18. Region importance – OASIS-1 dataset – HC vs AD



Supplementary Figure 8 - Regional deviations of the AD group from the OASIS-1 dataset. The marker indicates the mean effect size between the HC and the AD groups. The horizontal bars indicate the 95% confidence interval calculated using the percentile method on the bootstrap analysis.

19. Region importance – MIRIAD dataset – HC vs AD



Supplementary Figure 9 - Regional deviations of the AD group from the MIRIAD dataset. The marker indicates the mean effect size between the HC and the AD groups. The horizontal bars indicate the 95% confidence interval calculated using the percentile method on the bootstrap analysis.

20. Comparison of deviation values for healthy controls and patients - confidence interval of the differences

When we analysed the confidence interval of the difference between groups in the observed deviation, we obtained that, for the ADNI dataset, the difference between HC and EMCI was in the range [-0.03, 0.00], the difference between HC and LMCI was in the range [-0.06, -0.03], the difference between HC and AD group was in the interval of [-0.16, -0.06], the difference between EMCI and LMCI was in the range [-0.03, -0.02], the difference between EMCI and AD was in the range [-0.14, -0.06], and the difference between LMCI and AD was in the range [-0.10, -0.03]. For the AIBL dataset, the difference between HC and MCI was in the range [-0.09, -0.05], the difference between HC and AD was in the range [-0.17, -0.07], the difference between MCI and AD was in the range [-0.09, 0.00]. For the ARWiBo dataset, the difference between HC and MCI was in the range [-0.08, -0.03], the difference between HC and AD was in the range [-0.24, -0.10], the difference between MCI and AD was in the range [-0.16, -0.06]. For the OASIS-1 dataset, the difference between HC and AD was in the range [-0.18, -0.33]. Finally, for the MIRIAD dataset, the difference between HC and AD was in the range [-0.16, -0.41]. In summary, the five independent datasets presented mean deviation scores significantly different between their groups, except the comparison between HC and EMCI in the ADNI dataset and the comparison between MCI and AD in the AIBL dataset.

21. Traditional machine learning classification - confidence interval of the differences

To identify significant differences between the performance of the normative models and traditional classifiers, we calculated the confidence interval (95% of confidence) of the difference in AUC between the two methods. For the ADNI dataset we found that when classifying HC and EMCI the difference was in the range [-0.28, -0.09], when classifying HC and LMCI the difference was in the range [-0.24, -0.04], and when classifying HC and AD the difference was in the range [-0.25, -0.12]. For the AIBL dataset, when classifying the HC and MCI the difference of performance was in the range [-0.17, 0.74], and when classifying HC and AD the difference was in the range [-0.29, -0.12]. For the ARWiBo dataset, when classifying the HC and MCI the difference of performance was in the range [-0.15, 0.12], and when classifying HC and AD the difference was in the range [-0.17, 0.00]. For the OASIS-1 dataset, when classifying HC and AD the difference was in the range [-0.25, 0.04]. Finally, For the MIRIAD dataset, when classifying HC and AD the difference was in the range [-0.15, 0.06]. In summary, the traditional classifiers were superior to the normative models when predicting the difference between the groups in the ADNI dataset and the difference between HC and AD for the AIBL dataset; in contrast the performance of the two approaches was comparable for all other comparisons.